The Acoustic-to-Articulatory Mapping of Voiced and Fricated Speech

DISSERTATION

Presented in Partial Fulfillment of the Requirements for

the Degree Doctor of Philosophy in the

Graduate School of The Ohio State University

By

Edward L. Riegelsberger, M.S.

* * * * *

The Ohio State University

1997

Dissertation Committee:

Professor Ashok Krishnamurthy, Adviser

Professor Lee Potter

Professor Hitay Özbay

Professor Keith Johnson

Approved by

Ashok K. Krishnamurthy

Adviser

Department of Electrical Engineering

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	1997	Final

4. TITLE AND SUBTITLE
The Acoustic-to-Articulatory Mapping of Voiced and Fricated Speech

5. FUNDING NUMBERS

6. AUTHORS

Edward L. Riegelsberger

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

The Ohio State University

AFRL-SR-BL-TR-98-

0028

ON

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/NI
110 Duncan Avenue, Room B-115
Bolling Air Force Base, DC 20332-8080

NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT

Approved for Public Release

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*
See attached.

14. SUBJECT TERMS	15. NUMBER OF PAGES
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

ABSTRACT

Acoustic-to-articulatory mapping is the estimation of a time-varying vocal-tract shape from an acoustic waveform. While most research in acoustic-to-articulatory mapping considers only purely voiced speech, this dissertation investigates the problem for speech that includes fricatives. Aspects of fricative production and perception challenge many of the assumptions and techniques used in existing acoustic-to-articulatory mapping algorithms. This work investigates these issues and extends existing techniques for the acoustic-to-articulatory mapping of purely voiced speech to unvoiced and voiced fricatives in isolation and in continuous speech.

Linked-codebooks are used to examine the acoustic-to-articulatory mapping of voiced and unvoiced static fricatives. Acoustic-to-articulatory mapping performance is evaluated by analyzing articulatory estimation error for a number of synthetic fricatives and phonetic class clustering for a collection of real fricatives. Scatter plots of acoustic-to-articulatory mapping results on unvoiced fricatives demonstrate good phonetic class clustering and inter-class separability. Constraints on solutions are necessary to eliminate physically implausible solutions. For equivalent performance on voiced fricatives, the acoustic features had to be modified to deemphasize frequencies below 1 kHz.

Linked-codebook lookup, along with dynamic programming, is used to perform acoustic-to-articulatory mapping of continuous, purely voiced speech. Direct application of the algorithm to speech containing fricatives suggests that purely voiced acoustic-to-articulatory mapping provides contextual information that can improve fricative acoustic-to-articulatory mapping. The acoustic-to-articulatory mapping of intervocalic fricatives using different forms of contextual information demonstrates this point. A five step procedure is developed for the dynamic acoustic-to-articulatory mapping of continuous, voiced speech containing intervocalic fricatives. Multiple stages of processing are used to bootstrap articulatory estimates using contextual information. A collection of vowel-fricative-vowel tokens is used for development and testing. In most cases, the estimated articulatory trajectories appear natural and form fricatives with the correct place of articulation. Occasional errors occur due to vowel or fricative misidentification early in the optimization process. Problems in the vowel-fricative transition and source parameter optimizations ultimately limit the perceptual quality of the resynthesized speech.

To my wife, Whitney, for her unfailing love, support, and understanding.

ACKNOWLEDGMENTS

I am extremely grateful to my advisor and friend, Ashok Krishnamurthy, for his endless supply of helpful guidance, patience, and encouragement. Thanks go to the members of the IPS Lab, which has been like a family to me. Special thanks to Lee Potter, Stan Ahalt, Mike Collins, Bill Pierson, Mark Hanes, and Jayanth Anantharaman for their time and encouragement.

I would like to acknowledge the Air Force Office of Scientific Research (AFOSR), The Ohio State University, and the Department of Electrical Engineering for their financial support.

Finally, I would like to thank my wife, parents, and family who have stood by me during this long endeavor.

VITA

September 4, 1967	Born - Lima, Ohio, USA
1990	B.S. Electrical Engineering
1992	M.S. Electrical Engineering
1992-1995	Air Force Laboratory Graduate Fellow, The Ohio State University
1996-present	Graduate Teaching Associate, The Ohio State University

PUBLICATIONS

FIELDS OF STUDY

Major Field: Electrical Engineering

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Acoustic-to-articulatory mapping is the estimation of the vocal-tract shape during speech from only the acoustic waveform. As depicted in Figure 1.1, acoustic-to-articulatory mapping is the reverse of the speech production process, and is often referred to as *inverse mapping*. The shape of the vocal-tract, referred to as the vocal-tract *area-function*, is of great importance to linguistics and has direct applications to engineering and medicine. For engineering purposes, vocal-tract area-function data offer the potential for improved speech coding, synthesis, and recognition. It is generally accepted that articulatory representations can offer more parsimonious descriptors of speech than acoustic representations. This suggests the possibility of high quality, low bit-rate ($\leq$ 4 kbit/s) speech coding [1]. For similar reasons, the availability of area-function data would help the development of more accurate and efficient articulatory models, synthesis schemes, and control strategies for articulatory speech synthesizers. Some researchers suggest that speech recognition may be performed more effectively in an articulatory domain [2, 3], where the effects of coarticulation can be easily compensated. Acoustic-to-articulatory mapping solutions are necessary to make this a reality.

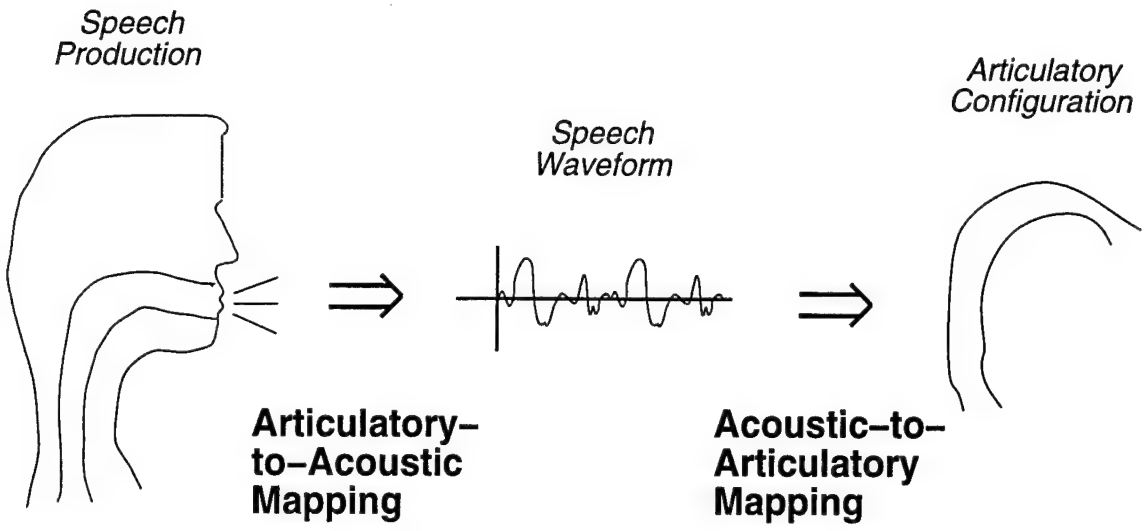


Figure 1.1: During speech production, vocal-tract shape is varied to produce a speech waveform. This speech production process is an articulatory-to-acoustic mapping. Its inverse, acoustic-to-articulatory mapping, attempts to estimate the vocal-tract shape from the speech waveform.

Acoustic-to-articulatory mapping is a challenging problem that has been studied for over 30 years [4]. Past work suggests that the lofty goal of area-function recovery is difficult to realize. Time-varying area-functions must be estimated from sampled speech which, due to its band limited nature, contains a limited amount of information. It is well documented that the acoustic-to-articulatory transformation is not one-to-one and that the presence of multiple valid solutions can complicate the acoustic-to-articulatory mapping problem. Analytic solutions for deriving area-functions from the speech waveform exist only for a few special cases. As a result, approaches to acoustic-to-articulatory mapping must take an analysis-by-synthesis approach that, based on some model of speech production, iteratively searches for the area-function whose synthetic speech best matches the original speech. Research in speech production and acoustics has produced a variety of articulatory-based speech synthesis

schemes which simulate the acoustics of sound production in the human speech mechanism [5, 6, 7, 8]. Of course, these synthesis systems can only approximate speech production in humans. A number of assumptions and simplifications must be applied to make the simulation tractable and to account for gaps in our knowledge of the speech production process. The assumptions and simplifications necessary for a tractable synthesis system can severely bias the results of analysis-by-synthesis, and limit the ability of inverse solutions to produce the true vocal-tract shape.

Work on the acoustic-to-articulatory mapping problem started with isolated vowels. In this *static* case, a single area-function is estimated from a single frame of speech. Attempts to solve the problem analytically discovered that the acoustic-to-articulatory transformation does not have a single unique solution. The few analytic techniques proposed [9, 10, 11] have unrealistic requirements such as acoustic measurements with infinite bandwidth, a vocal-tract without losses, and unnatural boundary conditions.

As computational power increased, analysis-by-synthesis approaches to the acoustic-to-articulatory mapping of static vowels such as [12] became feasible. Many techniques were proposed for avoiding local minima during optimization and dealing with the non-uniqueness problem. Acoustic-to-articulatory mapping systems for continuous, purely voiced speech [13, 14, 15] were also investigated. In this *dynamic* acoustic-to-articulatory mapping problem, requiring continuity in the time-varying articulatory shape was found to alleviate much of the non-uniqueness problem. A good survey of acoustic-to-articulatory mapping work can be found in [4]. A good review of past work can also be found in [16, 1].

Extension of these dynamic algorithms to include consonants has proven difficult. Consonant speech contains many characteristics that make acoustic-to-articulatory mapping more challenging than for purely voiced speech. Intervals of silence, such as during closure in stops, have a reduced information content which can make estimation more difficult. Downstream acoustic energy sources due to frication can be mixed with energy due to voicing, and must be separated. For many consonants, the vocal-tract shape can have a significant effect on the acoustic energy source which prevents the source and tract from being decoupled as is commonly done for voiced sounds. Forward models of consonants are not as well developed as for vowels due to the transient and sometimes non-linear nature of consonants.

Little has been reported on the acoustic-to-articulatory mapping of speech containing consonants. Acoustic-to-articulatory mapping of isolated unvoiced fricatives has been performed by Sorokin [17] and Shirai and Masaki [18]. Badin and Abry [19] report inverse mapping on continuous speech containing voiced fricatives. The acoustic-to-articulatory mapping system of Schroeter and Sondhi [1, 4] is formulated to include consonant speech, but few examples are given and unsatisfactory results are reported for fricatives.

The research described herein considers the extension of existing acoustic-to-articulatory mapping algorithms to consonant speech, specifically fricatives. Static fricatives are considered first using a table-based procedure known as linked-codebook lookup. Results show that fricative inversion is very susceptible to the many-to-one mapping problem, but that with sufficient constraints, reasonable and consistent results can be achieved. Acceptable performance for voiced fricatives requires the use of acoustic features that de-emphasize the influence of frequencies below 1 kHz.

Dynamic fricative acoustic-to-articulatory mapping is considered by combining the results for static fricatives into an algorithm for the inverse mapping of voiced, continuous speech. The ability of contextual information to improve fricative estimates is described. The capabilities and limitations of the algorithm are explored with vowel-fricative-vowel tokens from a male speaker.

This dissertation is organized as follows. The acoustic-to-articulatory mapping problem is formally defined in Chapter 2. The articulatory speech synthesizer used for our experiments is described in Chapter 3. In Chapter 4, linked-codebooks are used to study the static acoustic-to-articulatory mapping of vowels and fricatives using synthetic and real data. In Chapter 5, an acoustic-to-articulatory mapping algorithm is developed for continuous, voiced speech by using dynamic programming along with linked-codebook lookup to select articulatory trajectories. The results for static fricative inversion are used in Chapter 6 to augment the dynamic acoustic-to-articulatory mapping routine to include intervocalic fricatives. Chapter 7 reviews our results and conclusions and discusses directions for future work.

CHAPTER 2

OVERVIEW

2.1 Problem Definition

The acoustic-to-articulatory mapping problem is defined in terms of the forward *articulatory-to-acoustic mapping.* This forward mapping represents the human speech production mechanism and the acoustic and aerodynamic processes within that mechanism which create an acoustic waveform. The forward mapping, $\mathcal{S}$, takes a time-varying function, $\mathbf{x}$, describing the state and shape of the speech production system and produces a speech waveform, s.

$$s = \mathcal{S}(\mathbf{x}) \tag{2.1}$$

The objective of acoustic-to-articulatory mapping is to infer the true time-varying vocal-tract shape and state from the speech waveform. This is the inverse of the forward mapping.

$$\mathbf{x} = \mathcal{S}^{-1}(s) \tag{2.2}$$

Since the definition of acoustic-to-articulatory mapping starts from a forward articulatory-to-acoustic mapping, it is often referred to as an inverse problem or even *the* inverse problem in certain contexts. As for all inverse problems, the existence

and uniqueness of inverse solutions are important issues. For acoustic-to-articulatory mapping, if s is created by a human, then its corresponding vocal-tract state, $\mathbf{x}$, will exist; however, due to evidence of compensatory articulation, it is unclear whether $\mathbf{x}$ will always be unique.

In acoustic-to-articulatory mapping systems, an articulatory speech synthesizer is used to simulate the speech production process. Due to computational requirements and our limited knowledge about the speech production process, an articulatory speech synthesizer, represented by $\hat{\mathcal{S}}$, only approximates the true forward model, $\mathcal{S}$. Approximations and inaccuracies in the forward model cause error in our estimates of $\mathbf{x}$.

$$\hat{\mathbf{x}} = \hat{\mathcal{S}}^{-1}(s) \tag{2.3}$$

$$\mathbf{e} = \mathbf{x} - \hat{\mathbf{x}} \tag{2.4}$$

Quite often, the vocal-tract volume is represented in terms of an *articulatory model*, $\mathcal{A}$, which describes vocal-tract shape in terms of the position of articulators such as the tongue, lips, and velum. An articulatory model can be considered a front-end to an articulatory speech synthesizer, converting some parametric description of articulator position and/or trajectory, $\mathbf{p}$, into an vocal-tract description, $\mathbf{x}$, which drives an articulatory speech synthesizer.

$$\mathbf{x} = \mathcal{A}(\mathbf{p}) \tag{2.5}$$

This is depicted in Figure 2.1. Components of $\mathbf{p}$ correspond to vocal-tract and glottal-source parameters. The articulatory model generally represents vocal-tract shape with fewer parameters by applying knowledge about speech physiology to constrain

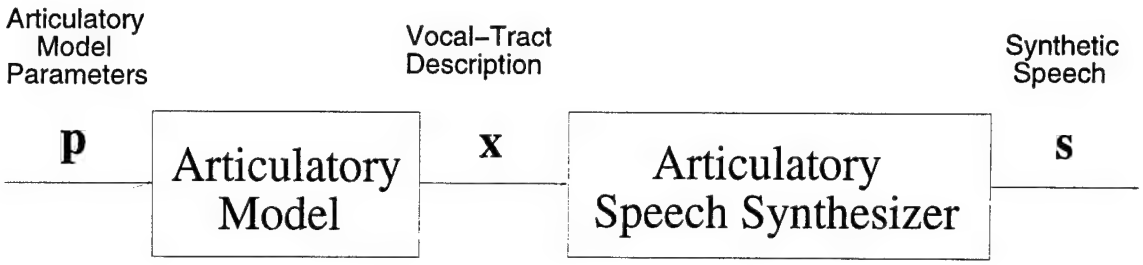


Figure 2.1: The forward articulatory-to-acoustic mapping.

vocal-tract shape and state to those physically realizable [20, 21]. The articulatory model is another source of approximation error in the forward model. The efficiency of an articulatory model description of the vocal-tract simplifies inverse mapping procedures and often outweighs the additional modeling error incurred.

In this new formulation of the inverse problem, the forward acoustic-to-articulatory mapping is approximated as

$$\hat{s} = \hat{\mathcal{S}}(\mathcal{A}(\mathbf{p})) = \mathcal{J}(\mathbf{p}) \tag{2.6}$$

and the corresponding inverse mapping is

$$\hat{\mathbf{p}} = \mathcal{A}^{-1}(\hat{\mathcal{S}}^{-1}(s)) = \mathcal{J}^{-1}(s). \tag{2.7}$$

The vocal-tract shape and state may then be calculated from the inverse solution.

$$\hat{\mathbf{x}} = \mathcal{A}(\mathcal{J}^{-1}(s)) \tag{2.8}$$

For most articulatory speech synthesizers there are no known analytic inverses, $\hat{\mathcal{S}}^{-1}$, available. A few analytic solutions exist for very simplified cases, but they require restrictive assumptions on the forward mapping, such as lossless acoustic wave propagation and precise knowledge of boundary conditions at the lips and the glottis [9, 10, 11]. Therefore, a common approach to calculating the inverse is called

analysis-by-synthesis. Let $D(s, \hat{s})$ be a measure of distance between two speech signals. Through efficient search procedures, an estimate, $\hat{\mathbf{x}}$, can be found that minimizes this distance measure.

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{y}} D(s, \hat{S}(\mathbf{y})) \tag{2.9}$$

Analysis-by-synthesis cannot guarantee the validity of a result if solutions are not unique. Analysis-by-synthesis techniques are also susceptible to local minima. Note that due to the "model mismatch" of the articulatory speech synthesizer, it is possible that a solution does not exist for which $D = 0$. In this case, the solution that minimizes D is chosen.

Acoustic-to-articulatory mapping is a challenging problem whose solution involves both the accurate modeling of the forward speech production process and the efficient estimation of the inverse of the forward model. Although inaccuracies in the forward model produce estimation error, some modeling assumptions are necessary and enable the inverse to be found without great difficulty. As research progresses in acoustic-to-articulatory mapping, the balance between accuracy and solvability will be further explored.

2.2 Research Issues in Acoustic-to-Articulatory Mapping

Ever since the first computer simulations of speech productions were performed, speech researchers have been searching for a way to determine, from acoustics alone, the dynamic shape of the vocal-tract during speech production. A large number of researchers have attempted the problem using a wide variety of approaches. Progress has been fueled by increasing computational power and better understanding of the

speech production process. While reasonable success has been achieved for voiced, non-obstruent, non-nasalized speech, the complete problem remains unsolved.

Any solution to the acoustic-to-articulatory mapping problem must address four interrelated issues. The wide variety of acoustic-to-articulatory mapping approaches can be understood and compared in terms of these issues.

1. **Which forward model should be used?** This choice is a tradeoff between physical accuracy and computational efficiency constrained by our incomplete understanding of the speech production process. The forward model selection has a significant effect on the other three research issues.

2. **What optimization techniques should be used?** While almost all approaches use analysis-by-synthesis, their implementation varies widely. Possible approaches include artificial neural networks (ANNs), table lookup, genetic algorithms, and brute-force optimization. Primary research issues include dealing with the non-uniqueness of solutions and avoiding local minima during optimization.

3. **What cost functions should be minimized?** All analysis-by-synthesis algorithms look for a solution that minimizes some *acoustic* distance measure. The type of acoustic distance measure affects whether acoustic-to-articulatory mapping solutions close in acoustic distance are close in articulatory distance. Distance measures also influence how well the resynthesized speech matches the original speech. While this third issue may be considered a part of the inverse mapping algorithm design, it is made a separate issue herein since it is a distinguishing feature of many acoustic-to-articulatory mapping attempts.

4. **How should results be evaluated?** Due to the assumptions and simplifications in the forward model and the limitations of analysis-by-synthesis techniques, acoustic-to-articulatory mapping estimates may not get very close to the true vocal-tract shape and state. This point, coupled with the fact that complete vocal-tract shape is unavailable for continuous speech, makes evaluating the results of acoustic-to-articulatory mapping difficult.

These four issues are explored in greater detail below.

2.2.1 Issue One: Selection of a Forward Model

Incomplete understanding of the speech production process combined with computational limitations prevents perfect modeling of the human speech production mechanism. Even the most complex articulatory speech synthesizers embody many assumptions and simplifications regarding speech acoustics, aerodynamics, and physiology. The accuracy of the forward model has a significant effect on the nature of acoustic-to-articulatory mapping algorithms and the quality of their solutions. Greater accuracy in the forward model generally improves the quality of inverse mapping solutions. As observed by Parthasarathy and Coker [14], *"automatic optimization methods work well only when the analysis synthesis model is capable of a very good fit, and the error can be driven substantially to zero. When the best choice of model parameters does not approximate the data well, known objective metrics do not make perceptually sensible compromises."* Unfortunately, synthesizers with greater accuracy and detail often require more elaborate input specifications. As more input parameters must be estimated from the same segment of speech, the estimation problem becomes harder. Additional information about speech production can help relieve the problem

somewhat by adding constraints to the solutions, but often this type of information is unavailable or incomplete. Therefore, selection of a forward model for acoustic-to-articulatory mapping is a tradeoff between modeling accuracy and computational complexity in both the forward model and the acoustic-to-articulatory mapping solution.

The wide variety of assumptions and design choices available for an articulatory speech synthesis will not be discussed here. Chapter 3 gives an overview of articulatory speech synthesis techniques and addresses some of the issues in selection of a forward model. One standard assumption will be discussed below, a common assumption that has a significant effect on the nature of acoustic-to-articulatory mapping solutions.

The vocal-tract forms a three dimensional volume through which acoustic waves propagate. While full three dimensional acoustic simulations exist, they are quite computationally demanding. Therefore a standard assumption in most articulatory speech synthesizers and the forward model of all known acoustic-to-articulatory mapping attempts is that of *planar* acoustic wave propagation. With only plane-wave propagation, the full three dimensional acoustic simulation reduces to a single spatial dimension. This significantly reduces computational complexity. Additionally, the vocal-tract representation reduces to a straight tube of non-uniform cross-sectional area. This is a more manageable description of vocal-tract shape that may be described by $x(l,t)$, a function of time, t, and location, l, along the total length, L, of the vocal-tract. This function of one spatial dimensional is commonly called an *area-function*. For convenience, many synthesizers use a one dimensional description

for all speech sounds, even though the assumption is not always valid. This form will be assumed throughout this document, unless otherwise noted.

Description of the vocal-tract via area-function, $x(l,t)$, does not utilize any information about speech physiology. The shape and range of motion of speech articulators such as the tongue, lips, jaw, and velum constrain the area-function in both shape and change of shape. Articulatory models attempt to represent the natural constraints imposed by articulatory kinetics. Articulatory models are the place where researchers have the greatest freedom to make changes in the forward model. Articulatory models are discussed in more detail in Chapter 3.

Most articulatory models depict the vocal-tract at a single instant in time. Although this frame-based approach does not take advantage of the slowly time-varying properties of vocal-tract articulators, it is by far the most common formulation. If parameters are estimated jointly over multiple frames, inter-frame correlation can be exploited through constraints on or parameterization of the articulatory trajectories. Parameterized articulatory trajectories can represent vocal-tract shape over many frames more efficiently and can help alleviate the many-to-one mapping problem, but are severely limited by the requirements of optimization [3, 14]. Optimization becomes less efficient as the number of parameters to be optimized increases. Commonly known as the "curse of dimensionality", feature size is a significant issue in articulatory-to-acoustic mapping. Feature size is a major issue in the selection of input and output representations, where a tradeoff must be made between modeling accuracy and the ability of an optimization algorithm to find a reasonable solution in a reasonable amount of time.

As an alternative to the frame or multi-frame-based descriptions, the kinematics of articulator motion can be represented as a dynamical system whose controlling inputs are then estimated. In this way, articulatory motion can be expressed by targets or goals which can be more easily linked to underlying phonemes or gestures [14, 16, 22]. These approaches offer great potential for solving the acoustic-to-articulatory mapping problem but require much more information about speech physiology and production in order to be implemented successfully.

Many representations have been proposed that represent articulation more efficiently than raw area-functions by reducing the dimensionality of the articulatory space and constraining the area-functions to those realizable by human articulation. Articulatory models allow for more natural description and interpolation of configurations than raw area-function representations. The choice of articulatory representation has a significant impact on optimization performance and must balance a number of tradeoffs. More parameters improve synthesis accuracy, but increase the dimensionality of the articulatory space making the optimization problem more difficult. Of course, there is the risk that modeling errors will over-constrain the problem, eliminating some potentially good solutions.

2.2.2 Issue Two: Techniques for Solution of Inverse Mapping

As discussed in Section 2.1, analytic solutions to the inverse mapping are unavailable for all but the simplest cases. As a result, numerical procedures must be used in an analysis-by-synthesis manner to solve the acoustic-to-articulatory mapping problem. Conventional multidimensional optimization procedures are iterative techniques, typically employing some form of gradient descent with constraints to find a locally

optimal solution. Non-gradient techniques, such as simulated annealing and genetic algorithms are also used.

While most acoustic-to-articulatory mapping efforts are alike in that they use analysis-by-synthesis, they widely differ in implementation. This is due, in part, to the variety of forward models used. Specifically, optimization for frame-based forward models differ from optimization for forward models based on dynamical systems. But for all approaches, two main problems must be addressed: local-minima, and non-uniqueness.

A major difficulty in the solution of the inverse problem for speech is the presence of multiple solutions. This non-uniqueness property has been well documented in the literature, both theoretically and empirically [23, 5]. Non-uniqueness manifests itself in a phenomenon known as compensatory articulation, where the acoustic effects of changes by one articulator are compensated by adjustments in other articulators. The most common example of compensatory articulation is the ventriloquist effect, in which a performer produces intelligible speech without any observable lip or jaw movement. Researchers have documented many compensatory relations. Hyoid height and lip protrusion can move together to keep the vocal-tract length unchanged. Rhotization of certain sounds, which lowers third formant frequency, can be produced by lip rounding and/or tongue raising in a compensatory manner. Other examples can be found in bite block experiments and lip-tug experiments.

The presence of non-uniqueness in the inverse mapping can also be shown from a theoretical point of view. Given a lossless vocal-tract of length, L, closed at one end and open at the other, the area-functions, $x(l), 0 < l < L$ and $1/x(L-l), 0 < l < L$,

produce the same transfer function. It is unclear how the presence of losses affects non-uniqueness.

When performing acoustic-to-articulatory mapping some of the effects of non-uniqueness can be reduced by constraining solutions only to those which are physiologically possible and by enforcing continuity over time in estimated vocal-tract trajectories. While some constraints can be imposed on the forward model through an articulatory model, other additional constraints can be incorporated into the error measure,

$$D = D_{acoust}(s, \hat{s}) + \rho \|\mathbf{p} - \mathbf{p}_0\|, \qquad \rho \geq 0. \tag{2.10}$$

The error measure consists of an acoustic distance, D_{acoust}, and an articulatory (geometric) distance weighted by ρ. The value, $\mathbf{p}_0$, can be defined as a neutral position or the previous frame estimate. Adding the geometric distance to the overall error metric helps to transform the optimization problem to a more well behaved one, much like regularization in the optimization of linear systems. This form of error measure has been used in many ways. Sorokin [24] added a static measure of muscle effort in the form of a distance from a neutral position to his optimization criterion. Schroeter, Larar, and Sondhi [13] used a geometric distance from the previous frame estimate in their error measure in which the weight, ρ, changed depending on the current speech class. Shirai and Kobayashi [3] used both distance from neutral position and distance from the previous frame estimate by adding a third weighted term to their error measure.

Another common problem in optimization is the presence of local minima in the error surface. Most numerical techniques risk converging to local minima resulting in non-optimal solutions. This is a second source of non-ideal solutions, separate

from the one of non-uniqueness, that adds to the number of possible solutions from which to choose. The problem of local minima is well documented in optimization literature and is difficult to avoid. Some new algorithms such as simulated annealing and genetic algorithms show potential for finding a global minimum in the limit.

One approach for reducing the local minima problem is to start the optimization process at the solution of the previous frame optimization, with the rationalization that articulatory configurations close in time should also be close in position. Of course, choosing a starting point for the first optimization remains a problem. This motivates an alternative method of selecting the starting point for optimization using linked codebooks or lookup tables [12, 15]. In these schemes, a table of articulatory shapes and corresponding acoustic consequences is generated. When choosing a starting point, the articulatory configuration whose consequence is closest to the desired one is used as the starting point. Schroeter and Sondhi [25] extended the codebook approach using dynamic programming. Instead of simply selecting the best starting point from a codebook according to some measure, the best starting *trajectory* is found by adding continuity constraints to the spectral measures in the dynamic programming scheme. This approach helps alleviate both the non-uniqueness and local-minima problems.

Artificial neural networks (ANNs) have also been considered for selecting starting points and even for tackling the entire acoustic-to-articulatory mapping problem. See [4] for a good survey of ANN techniques.

2.2.3 Issue Three: Definition of a Cost Function

To estimate $x(l,t)$ using the analysis-by-synthesis approach, a cost function, $D(s,\hat{s})$, between the original and resynthesized speech must be defined. This cost function, a measure of acoustic distance, should ideally be a good measure of $\mathbf{e}$, the distance between the estimated and actual vocal-tract shape. For good quality resynthesis, the cost function should also emphasize perceptually significant differences.

Many acoustic feature representations and distance metrics have been applied to the acoustic-to-articulatory mapping problem. Euclidean distance between (log) formant frequencies is very common [12, 16, 9, 10, 26, 27, 24] along with spectral/log-spectral distances [28, 29], and weighted cepstral distances [3, 14, 30, 31]. Some researchers have proposed using multiple error criteria [14, 31] in a sequential manner.

While the dimensionality of the acoustic space does not directly affect the optimization procedure, the acoustic representation can weight the importance of certain spectral differences over others. For example, Sorokin [24] found that optimizing the first three log formant frequencies gave better results than just two log formant frequencies, but four log formant frequencies gave worse results. One explanation for this observation is that adding additional acoustic constraints, in the form of additional formant frequencies, reduces the relative sensitivity to all the formant frequencies. Therefore, it is important that the acoustic vector weight perceptually significant differences over other differences.

Often, the acoustic distance of the cost function is augmented with other distance metrics that enforce continuity in the estimated articulatory parameters or apply some form of regularization. The design of these joint distance measures is a part of engineering the inverse mapping routine.

2.2.4 Issue Four: How to Evaluate Results

The way we evaluate the quality of acoustic-to-articulatory mapping solutions depends substantially on the motivation for pursuing acoustic-to-articulatory mapping. Ideally, a comparison between the estimated and actual vocal-tract shapes as in Equation 2.4 should be used. Accurate vocal-tract shape measurements are very difficult to obtain, so an absolute measure of estimation accuracy is generally unavailable. Recently, complete volumetric measurements of the vocal-tract during the production of steady state sounds have been obtained using magnetic resonance imaging (MRI) [32, 33]. With this data, accurate comparisons can be made. Unfortunately, these measurements are expensive and cannot yet be made on continuous speech. Partial articulatory measurements can be used for a qualitative evaluation of results. Possible measurements include x-ray photography, electropalatography, ultra-sonic imaging, or the tracking of pellets on articulators using electromagnetic or x-ray microbeam techniques.

Even if accurate measurements of vocal-tract shapes were available, current state of the art in acoustic-to-articulatory mapping is far from estimating true vocal-tract shape with much accuracy. Nevertheless, acoustic-to-articulatory mapping has many potential applications in speech recognition, coding, and synthesis. With these applications in mind, the quality of acoustic-to-articulatory mapping results can be measured indirectly in many ways. A very common measure of acoustic-to-articulatory mapping results is the quality of the resynthesized speech. This is the "speech mimic" approach to acoustic-to-articulatory mapping. Another way to evaluate acoustic-to-articulatory mapping results is to see how well they perform as features for speech recognition.

Sometimes, acoustic-to-articulatory mapping is applied to synthetic speech, for which the true vocal-tract shape in known. This is effective for testing solution techniques but avoids the serious "model mismatch" issue that plagues most acoustic-to-articulatory mapping systems.

2.3 Acoustic-to-Articulatory Mapping of Non-Voiced Speech

Most of the work in acoustic-to-articulatory mapping has been on voiced, non-obstruent and non-nazalized speech. One reason is that the forward models for voiced speech are the most accurate of all speech sounds. Additionally, perceptually-based cost functions available from research in speech recognition and speech coding are generally tailored to voiced sspeech. The speech production process for sounds such as nasals, fricatives, and stops is not as well understood. Overcoming these obstacles for the acoustic-to-articulatory mapping of consonants is a great challenge.

Acoustic-to-articulatory mapping of fricative consonants has been studied by a number of researchers. Sorokin [17] and Shirai [18] both looked at the problem for the static unvoiced fricative case and reported reasonable inverse mapping results. Acoustic-to-articulatory mapping of voiced or dynamic fricatives has not been considered in isolation. Consonant inversion has been reported in the complete acoustic-to-articulatory mapping system of Schroeter and Sondhi [1, 4] and formulated in Parthasarathy and Coker [14]. While inversion of some nasal and stop consonants is described by Schroeter and Sondhi [4], they report unsatisfactory results for fricatives [1]. An elaborate investigation into fricative production modeling and inverse mapping was performed by Scully et al. [34]. They were studying fricative production by simulating spoken VFV tokens along with other acoustic and aerodynamic measurements using a

articulatory speech synthesizer. Acoustic-to-articulatory mapping for dynamic sounds using parametric models was used along with hand tuning to reproduce the speech tokens in high detail. The Speech Maps effort [35] has driven research in acoustic-to-articulatory mapping that includes consonant speech. Beautemps et al. [36, 37] used acoustic and articulatory measurements to derive sagittal width to area-function rule and estimate area-functions for vowels and unvoiced fricatives. Badin and Abry [19] continue this work and report inverse mapping on continuous speech containing voiced fricatives. They use smoothed formant frequencies alone to estimate area-function and constrained constriction area during the fricative.

2.4 Philosophy

Area-function recovery is a very difficult and arguably unattainable goal [38, 39]. The transfer function of a lossless vocal-tract does not uniquely specify its area-function. Theoretical non-uniqueness has already been mentioned. Information is not available at all frequencies due to the approximately 6 dB/octave roll-off of the speech spectrum. Additionally, the assumption of plane wave propagation, a standard assumption, breaks down above frequencies around 4 or 5 kHz. To further confound the problem, uncertainty about the excitation makes separating the speech wave into source and transfer function components difficult.

The analysis-by-synthesis approach to acoustic-to-articulatory mapping, as defined in Equation (2.9) and applied in most research attempts, actually performs feature matching, a task quite different from area-function recovery. The goal of

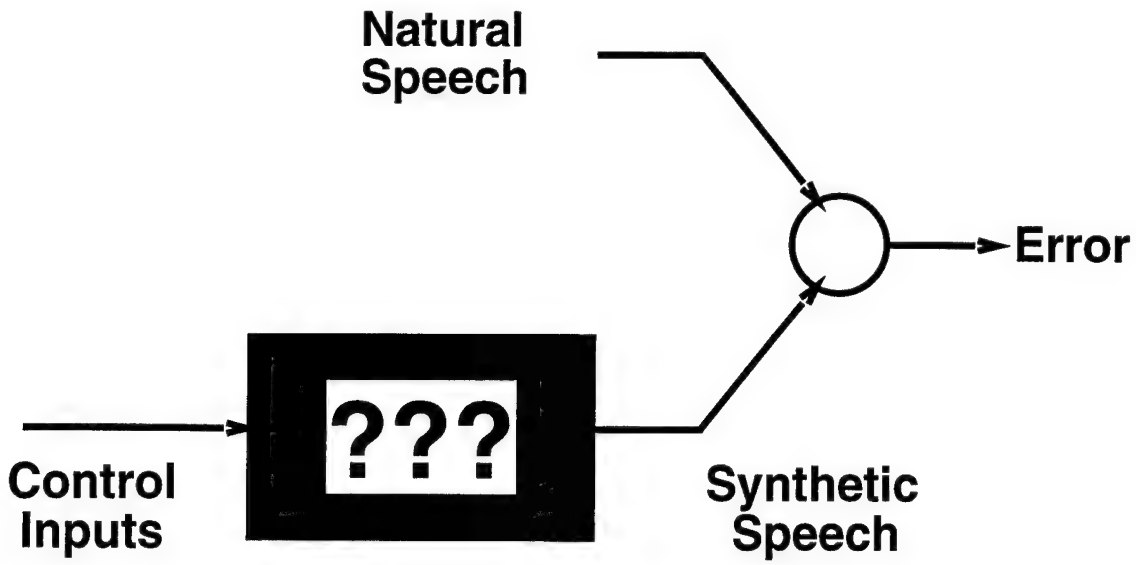


Figure 2.2: Black box viewpoint of acoustic-to-articulatory mapping.

area-function recovery supersedes that of feature matching; a good solution for area-function recovery will be a good solution for feature matching. The converse is not necessarily true unless the forward mapping is accurate.

The feature matching viewpoint can be generalized to an extreme by using the black box viewpoint of Figure 2.2. Given a forward model (black box), the goal of feature mapping is to estimate the control inputs to that model that minimize the error between the original utterance and the resynthesized version. The forward model can be anything: an articulatory speech synthesizer, a formant synthesizer, a LP coder, a modulation model, etc. Assuming that the black box is capable of reproducing all speech sounds, i.e. the range of the black box sufficiently spans the feature space, then a solution can always be found. All speech coding schemes can be represented in this way. Analysis-by-synthesis linear predictive coders (MPLP,

RPE, CELP, etc.) [40] are very similar in that they too use an analysis-by-synthesis approach to estimate some of their parameters.

So what are we trying to accomplish with acoustic-to-articulatory mapping if true area-function recovery is unrealistic, and what we really are doing is feature matching? Like speech coders, we attempt to derive control inputs to a forward model to match a given utterance as closely as possible. But by using a forward model resembling human speech production, estimated control inputs are an *articulatory representation* of speech.

Certainly, the closer our estimates are to the true area-function the better, but we assert that accurate area-function recovery is not necessary to produce useful articulatory representations. For example, articulatory phenomena such as coarticulation can be observed without requiring accurate area-function estimates. Information about place and manner of articulation is more significant than pure measurements such as the cross-sectional area 2 cm behind the velum. For many applications, preservation of the articulatory "gestures", without the unnecessary burden of complete area-function specification, will be sufficient.

What defines a good articulatory representation of speech? While we cannot yet answer this question, we suggest the following two properties as necessary conditions.

1. *An articulatory representation should vary in a slow, continuous manner.* This requirement is a natural consequence of speech physiology. Articulators have mass and inertia, and are moved by muscles of finite strength. Therefore, their motion should be continuous.

2. *An articulatory representation must be consistent in terms of place of articulation within and across different manners of articulation.* For example, constrictions for [s] should be located near constrictions for [n], [t], and [d]. Furthermore, constriction locations for alveolar sounds should be anterior of constriction locations for velar sounds.

The requirement of smoothly varying features is fundamental. Such features might be good for coding, allowing coarser sampling and better interpolation. Continuity is an even more significant requirement for consonants and makes articulatory representations a better descriptor for consonants. While many acoustic features of voiced sounds vary smoothly over time, no acoustic feature varies smoothly across obstruent consonants. Nevertheless, consonants are produced by smoothly moving articulators and therefore may be represented with smoothly varying articulatory representation.

The consistency requirement defines the way in which an articulatory representation must resemble true articulation. This resemblance is necessary in order to reveal many aspects of natural articulation that empower articulatory representations such as coarticulation and reduction.

The feature-mapping philosophy redefines the acoustic-to-articulatory mapping task; rather than estimating true vocal-tract state and shape, acoustic-to-articulatory mapping is generating an articulatory representation of speech. While the purpose of acoustic-to-articulatory mapping is different, the engineering problem of acoustic-to-articulatory mapping is essentially unchanged. The same four research issues must be addressed, although the use of a non-ideal forward model can now be justified. For our purposes, this viewpoint defines our stance on the fourth issue, evaluating results. The acoustic-to-articulatory mapping of vowels and fricatives will be based on

the fulfillment the above two properties and the production of intelligible resynthesis. While very qualitative in nature, these measures are realistic to the challenges of fricative inversion.

CHAPTER 3

ARTICULATORY SPEECH SYNTHESIS

Before the acoustic-to-articulatory mapping problem can be studied, a forward synthesis model must be selected. The characteristics of the forward model have a great influence on the nature of the inverse mapping and affect the form of acoustic-to-articulatory mapping solutions. Before the articulatory speech synthesizer used herein is described, the foundations of articulatory synthesis techniques will be discussed along with three general approaches to the problem, the assumptions and simplifications they make, and their effects on the inverse mapping problem.

3.1 Articulatory Synthesis Foundations

Articulatory synthesis techniques rely on many assumptions and simplifications to make the synthesis equations tractable and computable in reasonable time. As limiting as some of the assumptions are, reasonable synthesis performance still can be achieved.

The first, very reasonable assumption is that of linear acoustic wave propagation. Linear acoustic wave propagation in the vocal-tract is described by the equations

$$\frac{1}{\rho c^2}\frac{\partial p}{\partial t} + \nabla \cdot \mathbf{v} = 0 \tag{3.1}$$

and

$$\rho \frac{\partial \mathbf{v}}{\partial t} + \nabla p = 0, \tag{3.2}$$

where $p(x, y, z, t)$ is the sound pressure, $\mathbf{v}(x, y, z, t)$ is the particle velocity vector, ρ is the density of air, and c is the speed of sound in air. These three dimensional partial differential equations relate pressure and airflow. While it is possible to synthesize speech directly from Equations 3.1 and 3.2, the three dimensional simulation is extremely demanding computationally. Also, three dimensional information about the vocal-tract volume is not known accurately enough to utilize such precise calculations. If plane wave propagation is assumed, the vocal-tract becomes equivalent to straight tube of non-uniform cross-sectional area and the wave equation reduces to one dimension.

$$\frac{1}{\rho c^2} \frac{\partial (p(x,t)A(x,t))}{\partial t} + \frac{\partial u(x,t)}{\partial x} = \frac{\partial A(x,t)}{\partial t} \tag{3.3}$$

$$\rho \frac{\partial (u(x,t)/A(x,t))}{\partial t} + \frac{\partial p(x,t)}{\partial x} = 0 \tag{3.4}$$

$u(x,t)$ is the air volume velocity in the tube and $A(x,t)$ is the cross-sectional area of the tube. The plane wave assumption is valid as long as the acoustic wavelength is sufficiently larger than the cross-sectional dimension. This is generally true up to about 4kHz.

Losses are present in the vocal-tract due, in part, to viscosity, thermal conductivity, and wall vibration. Incorporating these losses into the one dimensional wave equations improves modeling accuracy. Although losses may be added into the one dimensional wave equations in many ways, those details will not be discussed here.

For computational purposes, the continuous one dimensional vocal-tract model must be sampled and discretized into a concatenation of tubes of differing cross-sectional area. Figure 3.1 depicts the simplification of the three dimensional vocal-tract volume to a discretized one dimensional tube model.

A number of articulatory synthesis techniques have been proposed for simulating the acoustic propagation of sound in the vocal-tract. The following three broad classes of articulatory synthesis techniques that use the above one dimensional linear plane wave propagation assumptions will be briefly described.

1. direct numerical evaluation of equations 3.3 and 3.4.

2. time domain simulation using wave-digital-filters (WDF), also known as the Kelly-Lochbaum model.

3. frequency domain simulation.

The first technique is the direct numerical evaluation of the wave equation. Runge–Kutta or a similar method is used to simulate the system. For stability the step size must be small which makes the simulation computationally intensive. Bocchieri [7] is a good example of this approach.

The second technique is also a time-domain technique and was first proposed for speech synthesis by Kelly and Lochbaum [41]. The technique uses a line analog model to approximate the vocal-tract as an transmission line. The Kelly–Lochbaum model, also known as a wave digital filter (WDF) or as a digital waveguide filter, is a discrete structure consisting of bidirectional delay lines (waveguides) and delay elements connecting adjacent sections [6]. It is a very popular choice for articulatory synthesis because of its speed. It can simulate time-varying area-functions but does

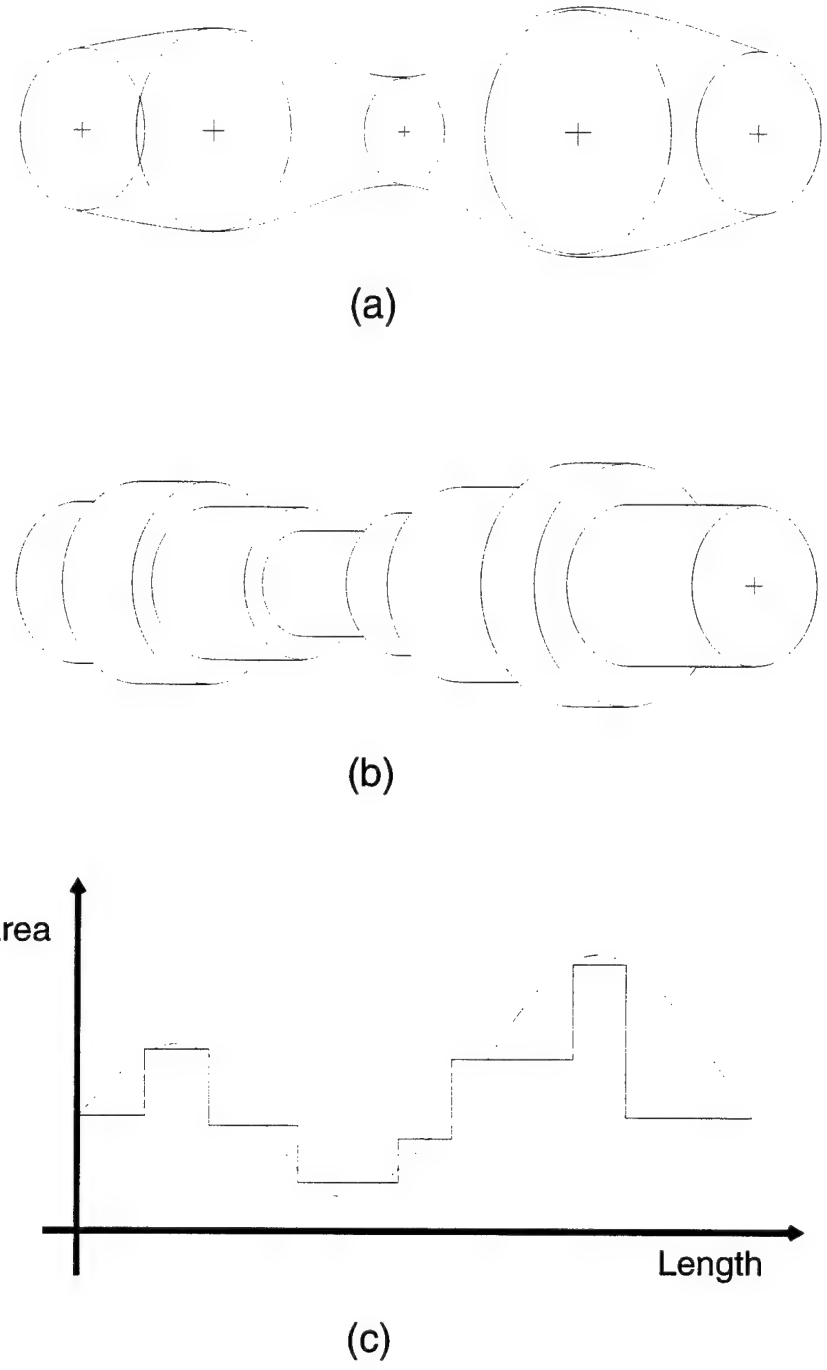


Figure 3.1: The simplifying reduction of a three dimensional vocal-tract volume to a discretized one dimensional tube model: (a) straight tube of non-uniform cross-sectional area, (b) discretized version of (a), (c) area-function of continuous and discrete tube models.

not allow frequency dependent losses. A drawback to the model is that it requires tube segments of a fixed length, causing continuous variation in vocal-tract length to be represented as discontinuous changes.

Frequency domain models can also be used for articulatory synthesis [5, 8]. As terminal analog models of the vocal-tract, they are easily expressed in terms of 2x2 "chain matrices". An advantage of the frequency domain approach is the convenience of modeling vocal-tract losses and radiation effects in the frequency domain. Frequency domain representations are much more convenient for acoustic-to-articulatory mapping since most of the features used for comparisons are based in the frequency domain.

Acoustic-to-articulatory mapping schemes could be developed for any of these synthesis schemes, but no one acoustic-to-articulatory mapping scheme would work for all synthesizers. The assumptions and associated drawbacks of each method require unique processing. For our the acoustic-to-articulatory mapping investigations, a frequency domain synthesizer will be used. The synthesizer is not the fastest in computation speed but it does not suffer from the discontinuities encountered in WDF synthesizers. Since optimization will be done with frequency domain speech features, the frequency domain synthesis approach is quite appropriate.

3.2 An Articulatory Speech Synthesizer

The articulatory speech synthesizer implementation used herein is a hybrid time-frequency domain synthesizer based on the synthesizer of Sondhi and Schroeter [8]. The hybrid synthesizer attempts to model aspects of speech production in their most natural form by having the vocal-tract represented in the frequency domain and the

glottal-source simulated in the time domain. The two representations are interfaced via an inverse Fourier transform and discrete convolution. The synthesizer includes a nasal branch with sinus cavities, coupled to the vocal-tract by a variable velum opening area. Fricative noise can be injected at the glottis for aspirated sounds or within the vocal-tract for fricative sounds. The glottal-source simulation uses a model of vocal-cord oscillation and is capable of reproducing many of the interactions between source and tract. The synthesizer is capable of producing all of the sounds of English.

The synthesizer implementation is written in C and uses signal processing and file management routines from Entropic Signal Processing System (ESPS) libraries. The synthesizer can produce detailed "monitor" files during synthesis that depict aspects of synthetic production including the source simulation, the fricative simulation, and transfer functions. The synthesizer can be driven by raw area-functions created using an area-function editor or by an articulatory model. Maeda's linear articulator model (LAM) [42] and Mermelstein's model [20, 43] have be implemented for this purpose. The synthesizer has been integrated into MATLAB using MATLAB's mex-file interface. The entire synthesizer, divided into a series of stages, may be accessed through MATLAB calls. This allows the state of the synthesizer to be examined and manipulated at any point. This is conducive to testing and prototyping new production models and fine tuning internal synthesizer parameters. An interactive graphic editor has been created using these MATLAB calls to assist hand synthesis of speech using the synthesizer.

The remaining discussion about the synthesizer reviews the primary aspects of its operation, but is not meant as a complete specification of its construction. Details

unique to the synthesizer or relevant to issues discussed elsewhere in this thesis, such as the fricative production model and constriction resistance, are described in detail. For a more thorough description of a hybrid time-frequency domain articulatory speech synthesizer implementation, see [8].

3.2.1 Acoustic Model

Each tube section in an acoustic tube model may be represented in the frequency domain as a two-port function described by a chain matrix.

$$\begin{bmatrix} P_{out}(\omega) \\ U_{out}(\omega) \end{bmatrix} = \begin{bmatrix} A(\omega) & B(\omega) \\ C(\omega) & D(\omega) \end{bmatrix} \begin{bmatrix} P_{in}(\omega) \\ U_{in}(\omega) \end{bmatrix} = K(\omega) \begin{bmatrix} P_{in}(\omega) \\ U_{in}(\omega) \end{bmatrix} \tag{3.5}$$

This chain matrix relates pressure, $P_{out}(\omega)$, and flow volume-velocity, $U_{out}(\omega)$, at the tube output to pressure, $P_{in}(\omega)$, and flow volume-velocity, $U_{in}(\omega)$, at the tube input. The elements, $\{A(\omega), B(\omega), C(\omega), D(\omega)\}$, are frequency domain quantities that incorporate vocal-tract losses. They are a function of the length and cross-sectional area of the tube section. Equations for the calculation of these variables along with some explanation of their derivation are available in [8]. For the remainder of this chapter, the frequency argument of these elements will be dropped for convenience. Figure 3.2 depicts the two-port representation of a tube section.

An N-tube model is described by the ordered product of the chain matrices of each tube section.

$$K_{Ntube} = \prod_{i=1}^{N} K_i = \begin{bmatrix} A_{Ntube} & B_{Ntube} \\ C_{Ntube} & D_{Ntube} \end{bmatrix} \tag{3.6}$$

From this chain matrix, the terminal characteristics of the entire tube model terminated by an impedance, Z_T, can be calculated. The tube model transfer function

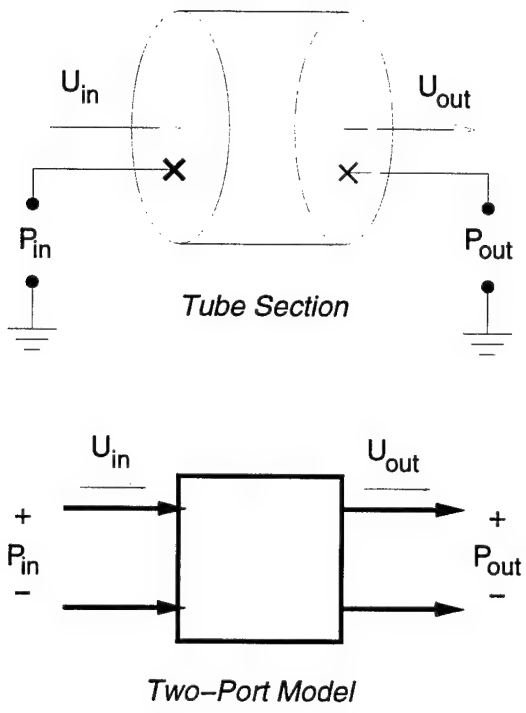


Figure 3.2: A tube section and its two-port functional representation.

is

$$H_{Ntube} = \frac{P_{out}}{U_{in}} = \frac{Z_T}{A_{Ntube} - C_{Ntube} Z_T} \tag{3.7}$$

and the impedance seen at the input of the tube model is

$$Z_{in} = \frac{D_{Ntube} Z_T - B_{Ntube}}{A_{Ntube} - C_{Ntube} Z_T}. \tag{3.8}$$

In a similar manner the transfer functions U_{out}/U_{in}, P_{out}/P_{in}, and U_{out}/P_{in} can be calculated.

Within the articulatory speech synthesizer, the functional description of the N-tube vocal-tract representation is the chain matrix, K_{tract}, which is calculated from the individual tube section chain matrices using Equation 3.6. While K_{tract} is sufficient to represent the vocal-tract filter, it is necessary to break the vocal-tract into a number of sections in order to couple the nasal-tract and insert frication. Figure 3.3 depicts the block circuit model assumed within the synthesizer. The vocal-tract is terminated with a lip radiation impedance, Z_{lip}, and is divided into four regions:

1. The pharyngeal region consists of tube sections between the glottis and the velum and is represented by the chain matrix, K_G. No closures are allowed in this region, i.e. the cross-sectional areas of each tube must be greater than zero.

2. The velar region consists of tube sections between the velum and the smallest constriction forward of the velum. If there is a constriction that produces frication, the velar region terminates at that constriction. If there is a complete closure in the vocal-tract, the velar region terminates at the closure. The chain matrix, K_C, describes the velar region.

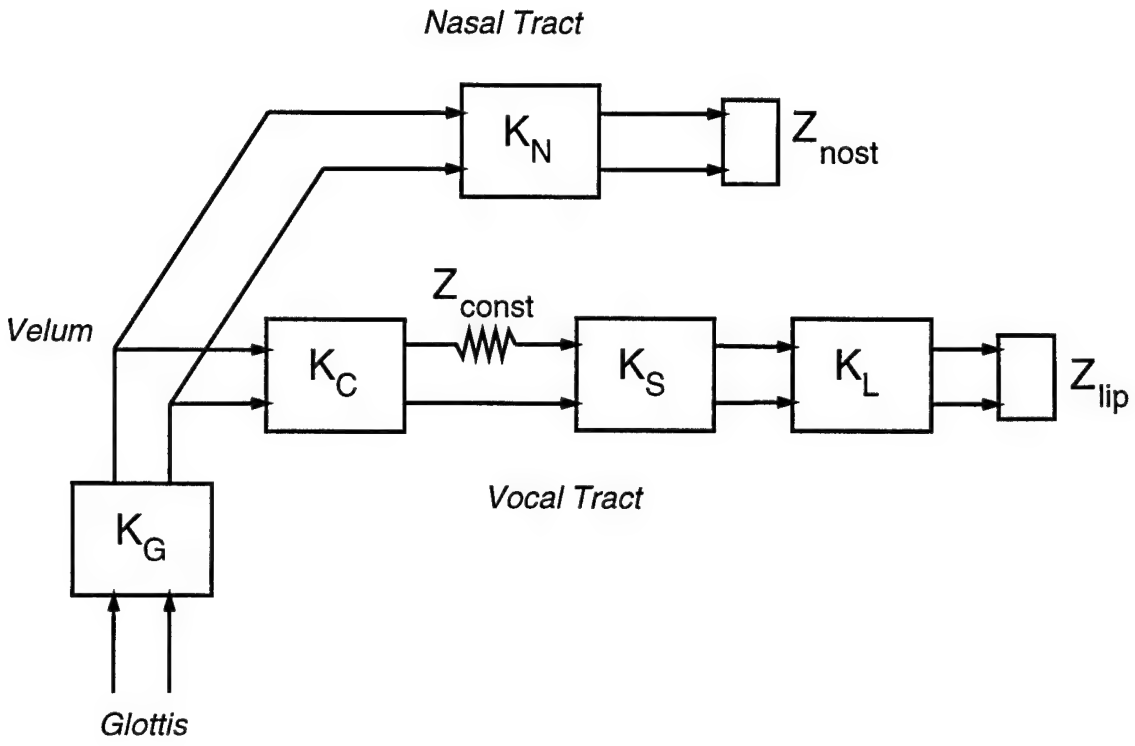


Figure 3.3: Acoustic model used in the articulatory speech synthesizer.

3. The fricative region consists of tube sections between the constriction and fricative noise source and is represented by the chain matrix, K_S. Even if frication will not be produced by the vocal-tract configuration, a frication noise source must be assumed somewhere between the constriction and the lips.

4. The forward region consists of tube sections between the fricative noise source and the lips. The chain matrix, K_L, describes this region.

The nasal-tract is also represented as a tube model terminated by a nostril radiation impedance, Z_{nost}. The describing chain matrix is K_N. While not depicted in Figure 3.3, the nasal branch also includes a Helmholtz resonator representing the sinus cavities whose effects are included in K_N. Coupling between the nasal-tract and vocal-tract is controlled by a velum area parameter.

The terminating impedances, Z_{nost} and Z_{lip}, represent the effects of radiation at the nostril and lips respectively. This radiation is modeled as that of a pulsating sphere with a radius equal to that of the opening as suggested by Flanagan [5].

The chain matrix from the glottis to the constriction is

$$K_{const} = K_{CR} K_C K_{cN} K_G \tag{3.9}$$

where K_{CR} is a special matrix that inserts a impedance, Z_{const}, at the constriction,

$$K_{CR} = \begin{bmatrix} 1 & -Z_{const} \\ 0 & 1 \end{bmatrix}, \tag{3.10}$$

and K_{cN} is a special matrix that accounts for coupling with the nasal branch,

$$K_{cN} = \begin{bmatrix} 1 & 0 \\ -1/Z_{in_{nasal}} & 1 \end{bmatrix}. \tag{3.11}$$

$Z_{in_{nasal}}$ is the input impedance to the nasal-tract seen at the velum. When the velum is closed, $Z_{in_{nasal}}$ is infinite and K_{cN} reduces to the identity matrix. The chain matrix

from the glottis to the lips is

$$K_{tract} = K_L K_S K_{const}. \tag{3.12}$$

The chain matrix from the glottis to the nostrils is

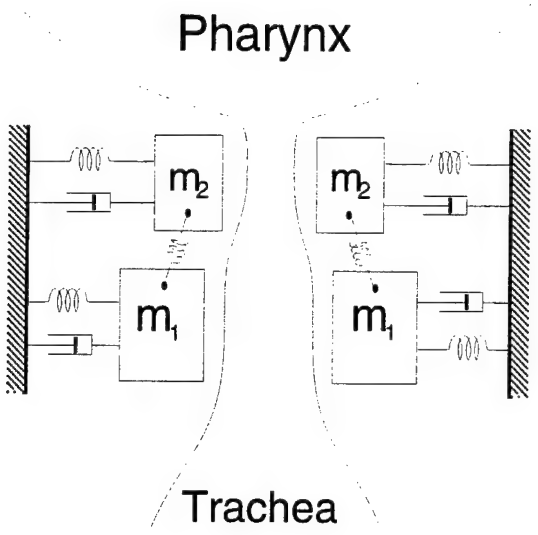
$$K_{nasal} = K_N K_{cV} K_G, \tag{3.13}$$

where K_{cV} is vocal-tract coupling matrix analogous to K_{cN}.

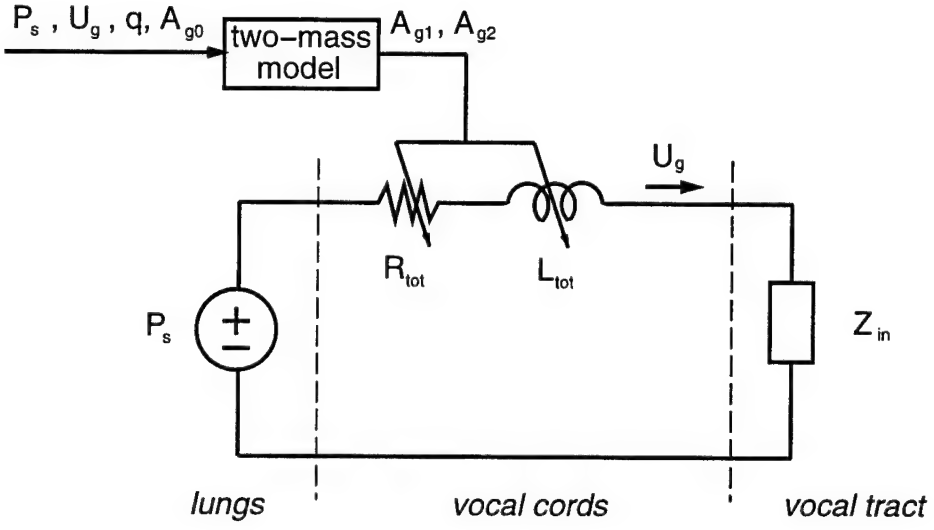
3.2.2 The Source Model

The glottal-source model represents the combined action of the lungs, trachea, and vocal-folds in regulating and modulating the flow of air into the vocal-tract. A variety of source models have been proposed with variations in the degree of physiological accuracy that range from modeling airflow in purely descriptive terms (F0, tilt, open quotient, etc.) to detailed physiological models of vocal-fold vibration and acoustomechanical coupling. The articulatory speech synthesizer presently employs the Ishizaka-Flanagan two-mass model [44], a model of some physiological detail. In this model, each vocal-fold is modeled by two coupled vibrating masses as in Figure 3.4(a). Symmetry is assumed between the two vocal-folds so that the motion of only one fold need be calculated. The time-varying positions of the two masses define the glottal cross-sectional area which affects the acoustic resistance and inductance of the glottal channel. Glottal volume velocity is calculated from the equivalent circuit diagram of Figure 3.4(b) given the vocal-tract input impedance, Z_{in}.

There are three time-varying input parameters to the two-mass model: lung pressure, P_s, which corresponds roughly with amplitude of voicing; glottal tension factor, q, which can be related to fundamental frequency; and glottal rest area, A_{g0}, which



(a)



(b)

Figure 3.4: Ishizaka-Flanagan two-mass mechanical model for vocal-fold motion and its equivalent circuit diagram for airflow through the glottis.

specifies the separation of the vocal-folds with no airflow and helps to control the onset and offset of voicing.

The glottal-source model produces a glottal volume velocity waveform, $u_g(t)$, which acts as the input to the vocal-tract filter, H_{VT}, and nasal-tract filter, H_N. H_{VT} and Z_{in} are calculated from the chain matrix K_{tract} using Equations 3.7 and 3.8, respectively. H_N is calculated from K_{nasal}. Synthesis output is produced by calculating the vocal-tract and nasal-tract transfer functions, adding them together, generating an impulse response function using the inverse Fourier transform, and convolving the impulse response with the glottal volume velocity waveform. Impulse responses are interpolated to produce smooth transitions between adjacent synthesis frames. Note that the vocal-tract and glottal-source models are continuous domain descriptions, therefore discretization is necessary in the implementation.

3.2.3 The Constriction Impedance

As the cross-sectional area of the constriction is decreased, the resistance to flow through the constriction is increased. The linear acoustic equations do not adequately model this effect when the constriction area gets small. Therefore, a discrete impedance, Z_{const}, has been inserted between the velar and fricative regions of Figure 3.3. This impedance is necessary to produce a pressure drop across the constriction which helps limit flow during fricatives and stops. The location of this impedance corresponds to the smallest constriction area in the vocal-tract. Only one constriction small enough to require this extra impedance term is assumed.

The constriction impedance includes a resistance and an inductance term representing losses at the contraction and expansion of the constriction. The constriction

impedance is a function of constriction cross-sectional area, A_c, length, L_c, and the rate of flow through the constriction, U_c.

$$R_{const} = \frac{\rho|U_c|}{2A_c^2} \tag{3.14}$$

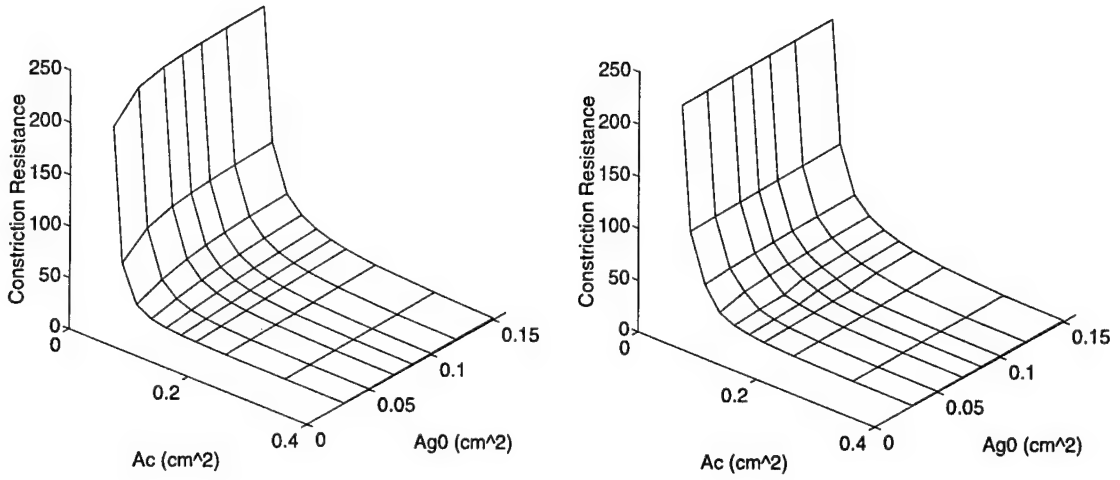
$$L_{const} = \frac{\rho L_c}{A_c} \tag{3.15}$$

Since the constriction resistance is flow dependent, it cannot be incorporated into acoustic transfer functions. Our solution is to approximate the constriction resistance, over each frame, by a fixed resistance that is a function of an (estimated) average flow for that frame. The differential equations of the two-mass model are too complicated to solve for average flow, so an approximate solution must be found. If we assume steady state flow, i.e., no vocal-fold motion, flow through the glottis will equal flow through the constriction. In this case, lung pressure, P_s, must equal the pressure drop across the glottis plus the pressure drop across the constriction,

$$P_s = U_g(R_g + R_{const}). \tag{3.16}$$

R_g, the resistance at the glottis, may be calculated directly from the two-mass model equations for glottal resistance (see Equation (8) of [44] or Equation (4) of [8]) by setting the glottal rest area of both masses equal and assuming no elastic expansion of the vocal-folds. By solving Equation 3.14, Equation 3.16, and the glottal resistance equation simultaneously, a value for flow may be found that matches, at least phenomenologically, the relation between constriction resistance, flow at the constriction, constriction area, and the glottal-source parameters q, A_{g0}, and P_s.

Figure 3.5(a) shows an example of the constriction resistance approximated using the above procedure as a function of constriction area and glottal rest area. This is

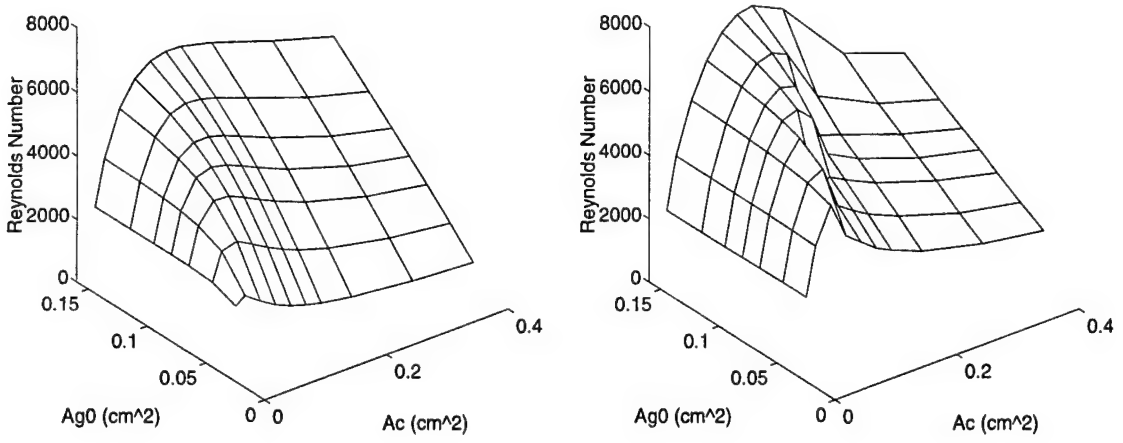


(a) Approximated constriction resistance

(b) Optimal constriction resistance

Figure 3.5: Comparison of approximated and optimal average constriction resistance for glottal tension, $q = 1$, and lung pressure, $P_s = 8$.

compared to the optimal values for R_{const} in Figure 3.5(b) obtained using an optimization procedure. Clearly, the approximated and optimal values are quite similar. The effect of approximation errors is better seen in the plots of Reynolds number at the constriction in Figure 3.6. Reynolds number is directly related to the amount of turbulence (frication) generated at the constriction. The high ridge in each plot is roughly the boundary between voicing and non-voicing, with voicing occurring for larger values of A_c. The higher, sharper ridge of the optimal Reynolds number plot results from the elastic expansion of the vocal-folds, which is not included in the approximation. Since we are already approximating a time-varying, flow-dependent resistance with a constant resistance, it is not known how significant accuracy is on the synthesis. If greater accuracy is needed, a table lookup/interpolation procedure



(a) Approximated Reynolds number

(b) Optimal Reynolds number

Figure 3.6: Comparison of approximated and optimal Reynolds number at the constriction for glottal tension, $q = 1$, and lung pressure, $P_s = 8$.

or artificial neural network can be used to more closely reproduce the values derived by optimization.

3.2.4 The Fricative Model

In addition to the synthesis of voiced sounds as described above, the synthesizer is capable of producing stop, fricative and aspirated consonants. For stops, the buildup and release of pressure during stop closures is accomplished by the buildup of supra-glottal pressure within the time-domain source simulation. For fricatives and bursts following stops, frication noise is injected at the appropriate location in the vocal-tract and is amplitude modulated by flow rate through the constriction. Aspiration is injected at the glottis much like frication is injected within the vocal-tract.

Frication noise is produced by airflow passing through a constriction in the vocal-tract downstream of the glottis. A sufficiently small constriction causes laminar flow to become a turbulent jet which acts as a noise pressure source. In some situations, the turbulent jet impinges on an obstacle downstream of the constriction such as the alveolar ridge, teeth, or lips, producing another noise pressure source with an amplitude greater than the one at the constriction. Within the synthesizer, only a single turbulence producing constriction is assumed, which must be located forward of the velum. The noise pressure source resulting from the constriction can be located anywhere between the constriction and the lips. The three sections within the oral-tract of the synthesizer as shown in Figure 3.3 are necessary so that the constriction and the frication noise pressure source can be independently located.

The frication pressure source is modeled as a random white Gaussian noise source whose amplitude is a function of airflow through the constriction. More specifically, its amplitude is related nonlinearly to airflow via the Reynolds number at the constriction as follows,

$$\begin{aligned} p_{fric}(t) &= G_{fric} random(t)(Re^2(t) - Re^2_{thresh}), & Re > Re_{thresh} \\ &= 0, & Re \leq Re_{thresh} \end{aligned} \tag{3.17}$$

where G_{fric} is an empirically determined gain term and *random* is a zero-mean, uncorrelated Gaussian random variable. Reynolds number is a dimensionless quantity related to the degree of turbulence in the flow of a fluid. Its squared value, Re^2, is defined as

$$Re^2(t) = \frac{4\rho^2}{\pi\mu^2}\frac{u_c(t)^2}{A_c}. \tag{3.18}$$

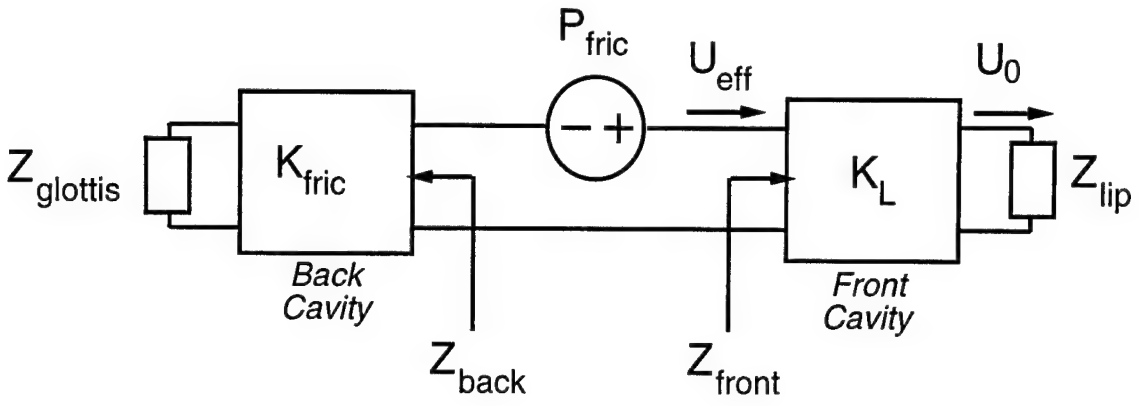


Figure 3.7: Block circuit diagram of a frequency domain fricative model.

where A_c is the cross-sectional area at the constriction and $u_c(t)$ is the flow through that constriction. Frication occurs only when air velocity through the constriction is above the threshold, Re^2_{thresh}.

Flow at the constriction is obtained by filtering glottal flow, u_g, with the transfer function, $H_U = U_c/U_g$, corresponding to the chain matrix K_{const}. Fricative output at the lips is produced by filtering the fricative noise pressure source waveform, p_{fric}, through the transfer function $H_F = P_{lips}/P_{fsource}$. The relation between the frication noise pressure, P_{fric}, and sound pressure at the lips, P_{lips}, is a function of the vocal-tract cavities both in front of and behind the pressure source as shown in Figure 3.7. The front cavity portion of the vocal-tract is from the frication source to the lips and has frequency characteristics described by the chain matrix K_L. The back cavity extends from the glottis to the frication source and is described by the chain matrix $K_{fric} = K_S K_{const}$. The presence of the back cavity produces zeros in the fricative spectrum. Given K_L and Z_{lip}, the front cavity transfer function $T_{fric} = U_0/U_{eff}$ and input impedance Z_{front} can be calculated using Equations 3.7 and 3.8. Given K_{fric}

and $Z_{glottis}$, which is tuned for no reflection at the glottis, Z_{back} can be calculated. The relation between the frication noise source and sound pressure at the lips is easily derived although it may be surprising that the poles and zeros of T_{fric} are not the poles and zeros of the fricative transfer function.

$$\frac{P_{lip}}{P_{fric}} = Z_{lip}\frac{U_0}{U_{eff}}\frac{U_{eff}}{P_{fric}} = \frac{T_{fric}Z_{lip}}{Z_{back} + Z_{front}} \tag{3.19}$$

3.2.5 Articulatory Models

While not a part of the acoustic synthesizer, articulatory models are an important part of articulatory speech synthesis. They model speech articulators such as the tongue, lips and jaw in order to generate reasonable area-functions using a small number of parameters.

A number of articulatory models depict the shape of the vocal-tract in the mid-sagittal plane based on x-ray photography. Two such articulatory models are presently available as a front-end to the articulatory speech synthesizer: Maeda's linear articulator model (LAM) [42] and Mermelstein's model [20, 43].

Figure 3.8 displays the linear articulator model. It models vocal-tract shape in the mid-sagittal plane using seven parameters that are based on a factor analysis of 400 x-ray profile images. The statistical analysis produced four components that can explain 98% of the variance in the observed profiles. Three of the four factors describe tongue shape and can be related to movements by the tongue body, the tongue dorsum, and the tongue apex. The fourth factor controls jaw angle. Three additional parameters control lip height, lip protrusion, and larynx height.

Figure 3.9 displays Mermelstein's model. It depicts vocal-tract shape in the mid-sagittal plane by modeling speech articulators such as the tongue, lips and jaw with

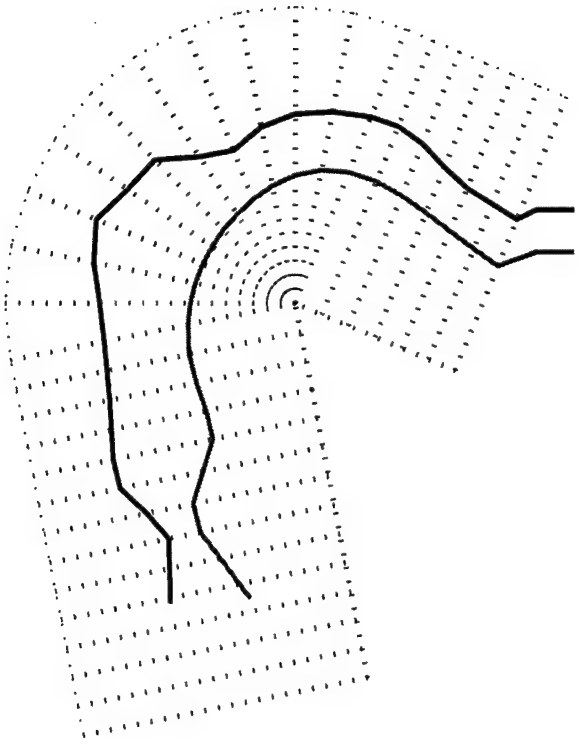


Figure 3.8: A vocal-tract configuration produced by the Maeda linear articulator model (LAM).

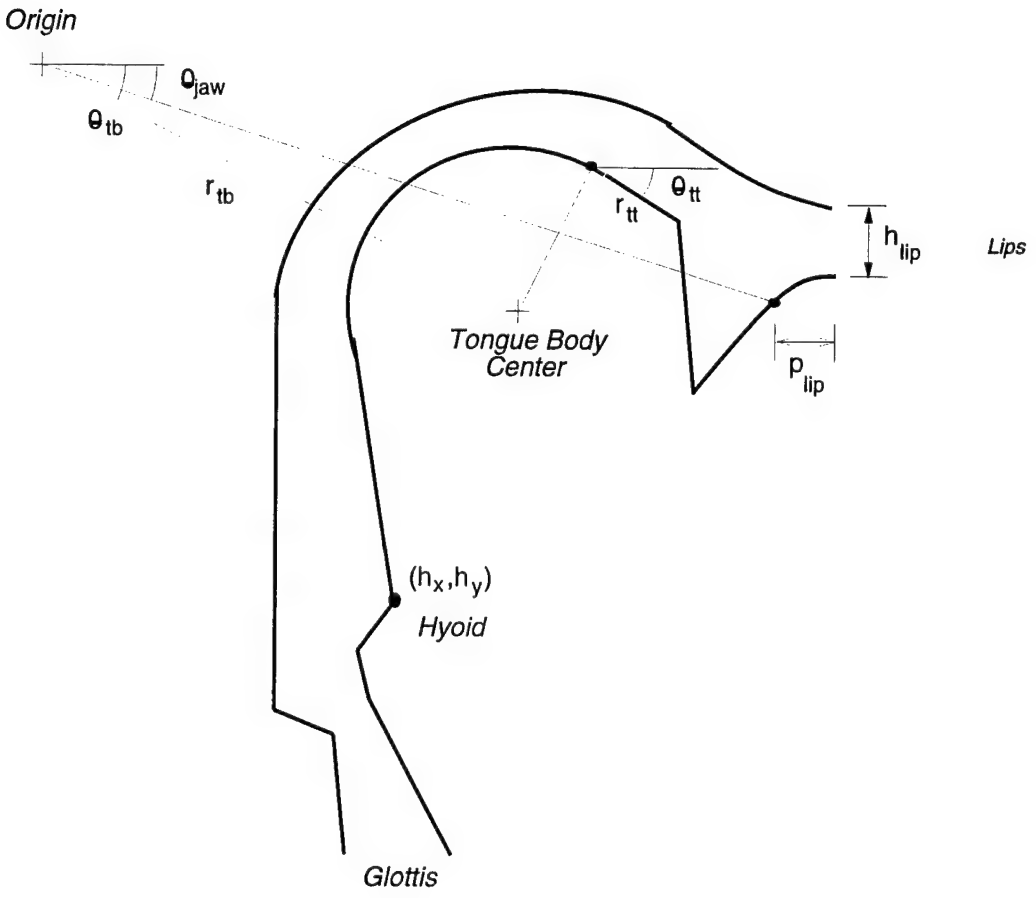


Figure 3.9: A vocal-tract configuration produced by the Mermelstein model.

geometric shapes. Mermelstein's model is controlled by ten parameters which specify hyoid position, (h_x, h_y), tongue body center in polar coordinates, (r_{tb}, θ_{tb}), tongue tip location relative to the tongue body in polar coordinates, (r_{tt}, θ_{tt}), jaw angle, θ_{jaw}, lip height, h_{lip}, lip protrusion, p_{lip}, and nasal coupling, A_{coup}.

For both models, area-functions are produced by converting mid-sagittal widths to cross-sectional areas using a mapping of the form

$$x(l) = \alpha(l) width(l)^{\beta(l)}, \tag{3.20}$$

where $\alpha(l)$ and $\beta(l)$ are fixed functions of location within the vocal-tract and are determined from physiological measurements.

Mermelstein's model has more freedom to model consonant configurations than the LAM. Unfortunately, extreme values of input parameters to Mermelstein's model can produce very unnatural looking configurations and occasionally cause the model to break down. It is difficult to predict which input parameters will produce these bad configurations. Limiting the range of individual parameters to avoid these configurations overly restricts the model and prevents many valid configurations from being produced.

CHAPTER 4

STUDY OF THE INVERSE TRANSFORMATION FOR VOWELS AND FRICATIVES

4.1 Introduction

The successful recovery of articulation from acoustics requires some understanding of the acoustic-to-articulatory transformation. By investigating the many-to-one mapping problem, the effects of model mismatch, and the avoidance of local-minima in optimization, we can identify reasonable heuristics for inverse mapping and get some idea of the potential accuracy of acoustic-to-articulatory mapping solutions.

Much can be learned by studying the acoustic-to-articulatory transformation for a single frame of speech. In this *static* problem, a single articulatory configuration is estimated from a segment of (presumably) steady-state speech. The more general *dynamic* problem of acoustic-to-articulatory mapping requires the estimation of an articulatory trajectory and is considered in Chapters 5 and 6.

This chapter, as well as the remaining chapters, will use a restricted definition of voiced speech that refers to only vowels and glides, without liquids or any fricated, aspirated, nasalized, or plosive sounds. Most existing acoustic-to-articulatory mapping algorithms are restricted to this limited class of voiced speech. The restricted

definition simplifies the distinction between fricated speech, which can be voiced, and voiced speech, which does not include fricatives.

In this chapter, we describe a technique known as linked-codebooks for the acoustic-to-articulatory mapping of static sounds. We then use this procedure to study the inverse transformation. Since fricative inversion is the primary thrust of this thesis, special attention will be given to fricative linked-codebooks and fricative acoustic-to-articulatory mapping. The results will be used to motivate the approaches taken in Chapters 5 and 6 for building dynamic acoustic-to-articulatory mapping systems for utterance containing both voiced and fricated sounds.

4.2 Linked-Codebooks

In the *static* formulation of the acoustic-to-articulatory mapping problem, we wish to infer an articulatory configuration from a single frame of speech. As described in Chapter 2, an analysis-by-synthesis approach is commonly taken to find the articulatory configuration whose acoustic result best matches the original speech. This may be formulated as a multidimensional, constrained optimization problem. In all optimization problems, the avoidance of sub-optimal solutions, or local minima, is a significant issue. Many researchers have reported that optimization techniques have a better chance of finding a global (or near optimal) solution if they start close to that solution. One approach for reducing the local minima problem is to start the optimization process at the optimized solution of the previous frames, with the rationalization that articulatory configurations close in time should also be close in position. Of course, choosing a starting point for the first optimization remains a problem. This

motivates the following rather popular and successful way of selecting the starting point for optimization using lookup tables called linked-codebooks [12, 15].

Let Φ be a C entry linked-codebook consisting of articulatory vectors, $\mathbf{p}_n, n \in [1, C]$, and their corresponding acoustic consequences, $\mathbf{q}_{feature,n}$, where $feature$ is a label identifying the type of acoustic feature representation used. More than one acoustic feature may be linked to the same articulatory vector. Whenever possible, the $feature$ subscript will be dropped for convenience. Notationally, the articulatory vector, $\mathbf{p}_n$, and the acoustic features, $\mathbf{q}_n$, of the nth linked-codebook entry, or codeword, are accessed as follows.

$$\mathbf{p}_n = \Phi_*(n) \tag{4.1}$$

$$\mathbf{q}_n = \Phi_{feature}(n) \tag{4.2}$$

Codebook lookup entails finding the codeword whose acoustic feature best matches that of the speech waveform, s, according to some distance measure, $D(s, \Phi_{feature}(n))$. The codebook index of the winning codeword is

$$n^* = \arg\min_{\mathbf{n}} D(s, \Phi_{feature}(n)) \tag{4.3}$$

and the best articulatory configuration is $\Phi_*(n^*)$.

Linked-codebooks are used to seed optimization in inverse mapping schemes. Using codebooks in this manner has a two-fold purpose. First, it helps alleviate the problems of local minima in optimization by (hopefully) starting the optimization close to the correct solution. Secondly, linked-codebooks can provide the N best fits, which, with continuity constraints or dynamic programming, can be used to eliminate some of the unreasonable non-unique solutions. For our analysis purposes, the

linked-codebook may be thought of as a very coarse approximation of the inverse function.

4.2.1 Linked-Codebook Generation

To generate a linked-codebook, the parameter space of the articulatory model must be sampled in some logical fashion. Sampling is followed by pruning to remove undesirable configurations, such as ones with complete closure, and to get reasonable coverage of the articulatory and acoustic spaces with fewer codewords. Finally, acoustic features are calculated for each configuration. Unless otherwise noted, the acoustic features extracted are a function only of the vocal-tract shape, so linked-codebook sampling does not include parameters controlling the glottal-source. While source-tract interaction exists and is modeled within the synthesizer, its effect in voiced sounds is ignored for practical reasons. Source-tract interaction for fricated sounds is more significant, so some fricative codebooks may require sampling of source parameters.

In sampling the articulatory space, we want to get the most accurate representation of the acoustic-to-articulatory transformation possible with a finite number of samples. This is quite a challenge given the high dimensionality of the articulatory space and relatively small number of samples that can be practically stored and retrieved. Figure 4.1 displays the number of samples required to sample an articulatory space of N dimensions on a uniform grid with k samples per dimension. Clearly, as N increases, the number of samples necessary to keep the same sampling density increases exponentially. For example, the LAM, a rather simplistic model of merely seven dimensions, requires over 16000 samples to produce a sampling density

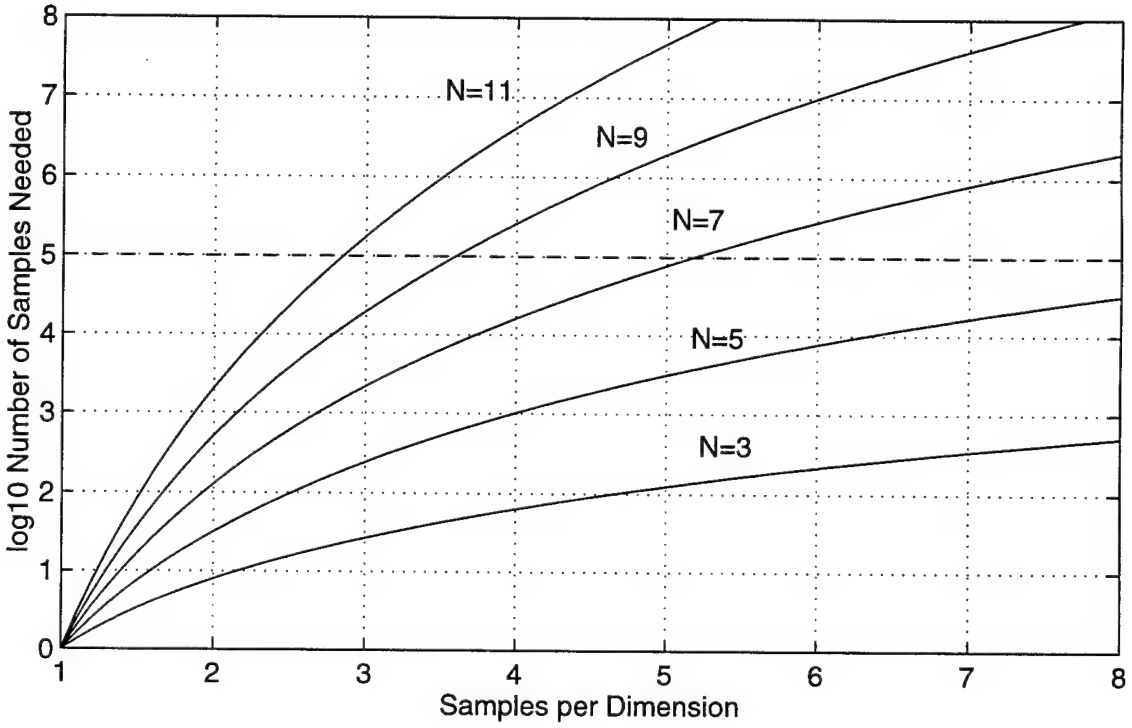


Figure 4.1: The number of samples required to sample on a uniform grid an articulatory space of N dimensions as a function of the number of samples per dimension. The dashed line at 10^5 samples indicates the upper limit on practical codebook size.

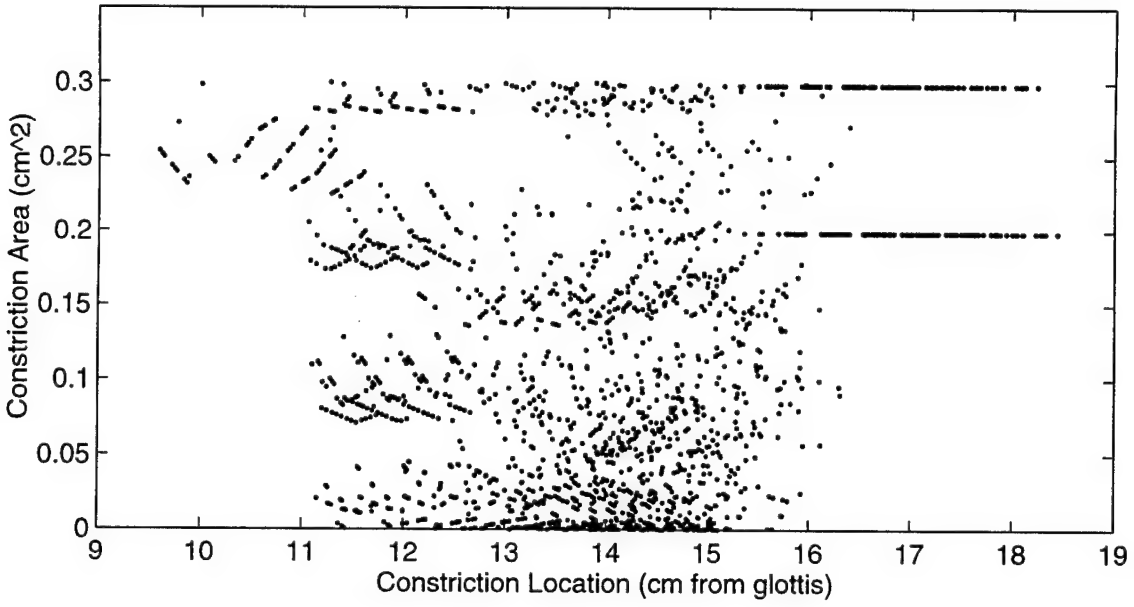
of 4 samples per dimension. Mermelstein's model, a slightly more detailed model with nine dimensions, requires over 260000 samples to achieve the same sampling density. Therefore, while more detailed articulatory models provide us with greater freedom and accuracy in controlling the vocal-tract shape, their increased dimensionality limits the effectiveness of acoustic-to-articulatory mapping procedures, including linked-codebook lookup.

The maximum allowable size of the linked-codebook is limited by three issues: codebook access (lookup) time, creation time, and memory consumption. Since this study is exploratory in nature, access time and creation time are less significant and the size of the codebook in memory becomes the dominant constraint. By keeping the size of linked-codebooks less than 5 Mb, and using an efficient method of encoding articulatory configurations and acoustic features, any codebook with fewer than 100000 entries is reasonable in size.

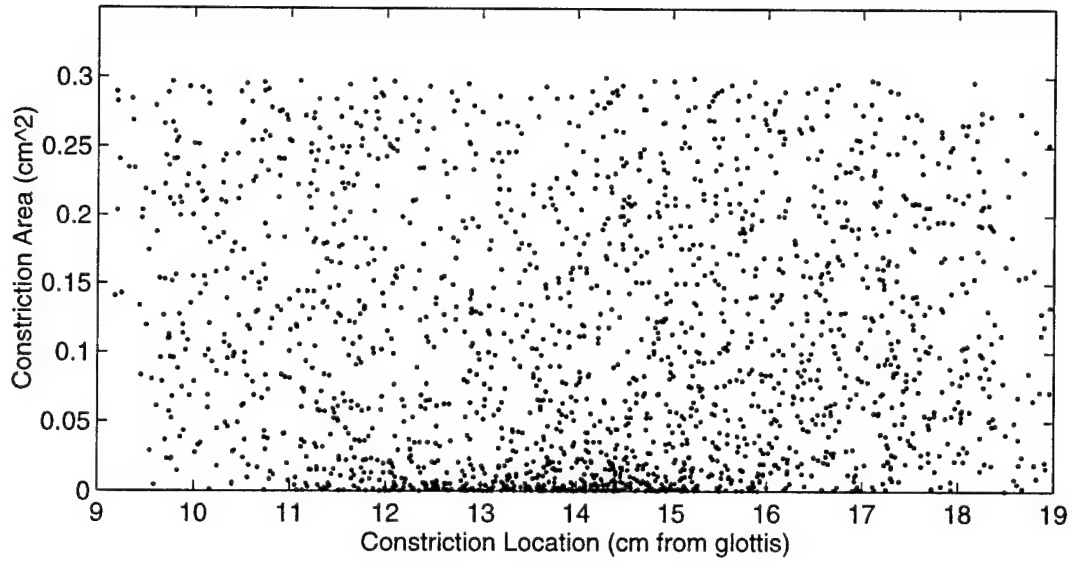
With such a relatively small number of samples to deal with, it is important to make sure that all regions of the articulatory space get adequate representation and that the codewords are used efficiently. This is accomplished with careful sampling, classification, and pruning. Sampling occurs over the region defined by the range of reasonable values for each parameter of the articulatory model. A number of strategies can be taken for adequately sampling this region. Sampling on a uniform (and logarithmic) grid was used by Atal [12] to generate a lookup table for vowels using a four parameter articulatory model. Random sampling [45, 39], typically uniformly distributed, is a popular alternative that appears to have an advantage, for higher articulatory dimensions, over sampling on a grid. For example, two articulatory codebooks with roughly 2000 entries each were generated: one by sampling on a

uniform grid with four samples per dimension, and the other by uniformly-distributed random sampling. Figure 4.2 shows a scatter plot for each codebook of the minimum constriction area and its location for each sample. By sampling uniformly on a grid, the full range of constriction locations and areas are not well represented. Random sampling does a much better job of evenly covering the articulatory space. Another approach to sampling the articulatory space taken by Larar et al. [46] is to sample paths connecting a set of reasonable "root" configurations to generate their codebook entries. With this "root-shape interpolation" method of sampling, configurations that are more likely to be used in real speech are more densely sampled, and unnatural or improbable regions of the articulatory space are avoided. The difficulty of this approach is in identifying a set of reasonable configurations that sufficiently cover the relevant articulatory space and don't leave any "holes".

For many articulatory models, including the Mermelstein model and the Maeda LAM, it is difficult to determine beforehand the nature of the vocal-tract shape for a given set of articulatory model parameter values. Therefore, each sample must be classified as a vowel, fricative, closed, non-English, or invalid configuration by examining the area-function it produces. Invalid configurations are those that produce an error condition within the articulatory model and correspond to physiologically impossible configurations. Closed configurations are any configuration with at least one complete closure (zero cross-sectional area) in the vocal-tract. The remaining three classes are distinguished by the location and size of the smallest constriction in the vocal-tract. Configurations with a minimum cross-sectional area greater than some threshold, A_{vowel}, are considered vowels. Configurations with a minimum cross-sectional area less than some threshold, A_{fric}, located anterior to the velum are



(a) Fricative codebook constrictions (4pt grid sampling).



(b) Fricative codebook constrictions (random sampling).

Figure 4.2: Scatter plot of constriction area and location for codebooks sampled randomly and on a uniform grid.

considered fricatives. The synthesizer, in its current form, can only produce frication for constrictions in front of the velum. Therefore, all configurations with a minimum cross-sectional area less than A_{vowel}, located behind the velum are considered invalid. This restriction is not a limitation for many languages, such as English, which do not use uvular or pharyngeal obstruents.

There is some ambiguity in the definition of a vowel or fricative configuration. Depending on the amount of airflow through the constriction, the same configuration could be perceived as a vowel or a fricative. That is why two separate thresholds, A_{vowel} and A_{fric}, are provided. Generally, A_{vowel} is less than A_{fric} allowing for some overlap in vowel and fricative definitions. Physical measurements of constriction area in talkers is difficult to obtain. Based on reasonable values for pressure and flow at the constriction during frication, Stevens [47] estimated constriction areas in the range of 0.1–0.2 cm^2. Stevens also points out that high vowels, such as [i] and [u], can have constriction areas less than 0.3 cm^2. Similar estimates by Badin et al. [48], but with more careful measurements of flow and pressure drop across the constriction, found constriction areas of 0.13 cm^2, 0.08 cm^2, and 0.15 cm^2 for [f,s,ʃ] respectively. Badin et al. also measured constriction areas from electropalatographic (EPG) data, estimating constriction areas of 0.02–0.04 cm^2 for [s] and 0.07–0.2 cm^2 for [ʃ]. An explanation for the discrepancy between the two sets of measurements is unavailable. Narayanan et al. [33] used magnetic resonance imaging (MRI) to measure the "static" 3-D geometry of four talkers producing eight English fricatives. They measured constriction areas as small as 0.098 cm^2 and as large as 0.299 cm^2. The above estimates give us a reasonable idea of the boundary between vowels and fricatives. Therefore, 0.2 cm^2 will be considered the boundary between vowels and fricatives. For

Class	Maeda		Mermelstein	
Total	84693	(100.0%)	176706	(100.0%)
Vowels	35088	(41.4%)	34739	(19.7%)
Fricatives	9761	(11.5%)	8199	(4.6%)
Both Vowel & Fricative	4912	(5.8%)	5261	(3.0%)
Closed	33657	(39.7%)	108433	(61.4%)
Non-English	1275	(1.5%)	12636	(7.1%)
Invalid	0	(0.0%)	7438	(4.2%)

Table 4.1: Classification of samples after uniformly distributed random sampling of the LAM and Mermelstein model articulatory spaces to generate 40000 entry vowel codebooks. ($A_{vowel} = A_{invalid} = 0.15\,\mathrm{cm}^2$ and $A_{fric} = 0.3\,\mathrm{cm}^2$).

purposes of resolving ambiguity and overlap between vowels and fricatives, A_{vowel} will be set to $0.15\,\mathrm{cm}^2$ and A_{fric} set to $0.3\,\mathrm{cm}^2$. Some additional tuning may be necessary based on the threshold of frication for the synthesizer.

In generating a vowel codebook of 40000 samples, the articulatory space of the Maeda LAM was sampled 84693 times. Table 4.1 shows the the number of configurations classified in each category along with the percentage of the total. Classification was performed with $A_{vowel} = A_{invalid} = 0.15\,\mathrm{cm}^2$ and $A_{fric} = 0.3\,\mathrm{cm}^2$. Clearly, closed configurations are a significant portion of the total number of samples. This suggests that the range of model parameters over which we sampled may be larger than necessary. But the set of non-closed configurations is not necessarily a convex one, therefore, it is probably best to overestimate the parameter range. Computationally, sampling and classification of 100000 samples takes about 3 hours on a 60 MHz Pentium. Therefore, we can accept low percentages for classes such as fricatives, and still generate rather large codebooks in about a day. It is interesting to note that for the Maeda LAM, approximately 29.5% of the non-closed configurations are capable

of producing frication. For comparison purposes, Table 4.1 shows similar statistics for generating another vowel codebook of the same size using Mermelstein's model. Mermelstein's model is more sensitive to extreme values of its input parameter values and, therefore, more configurations caused errors and were labeled invalid. Of the non-closed configurations, 27.9% are capable of producing frication.

A number of procedures have been proposed to produce a more efficient coverage of the acoustic-to-articulatory mapping by the codebook using clustering or pruning [12, 26, 46]. A procedure by Schroeter et al. [45] offers an effective and straightforward approach to improve codebook efficiency and coverage. The first three formants of each potential codeword are mapped onto a logarithmically spaced 3-D grid of formant frequency bins. Potential codewords are accepted into the codebook only if there are no geometrically similar codewords in the same formant frequency bin. Articulatory model configurations X and Y are geometrically similar if the distance

$$d_{geo} = \frac{1}{N}\sum_{i=1}^{N}(X_i - Y_i)^2 \tag{4.4}$$

is less than some threshold.

After sampling, classification, and pruning, we should have a reasonable representation of the articulatory side of the acoustic-to-articulatory mapping. The linked-codebook is completed by generating for each codeword, the corresponding acoustic features produced by that codeword's articulatory configuration. The choice of acoustic feature has a significant effect on the codebook's ability to perform an effective acoustic-to-articulatory mapping. A good feature is sensitive to perceptually significant differences, yet insensitive to glottal variation and speaker dependent differences. The type of acoustic feature used depends on the algorithm used and the

type of speech being analyzed. Therefore, the issue of acoustic feature selection will be further addressed as it arises in later sections.

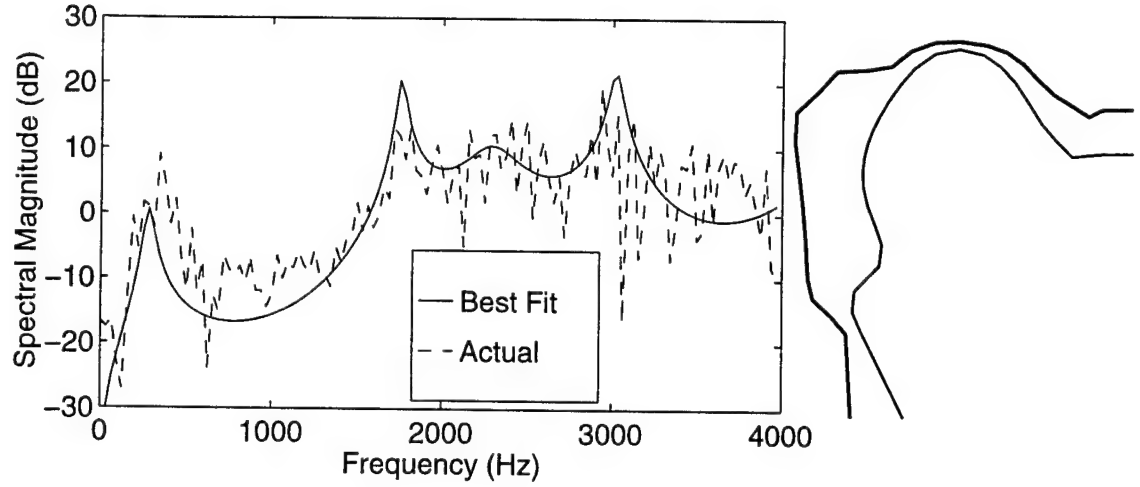
4.3 Linked-Codebook Lookup for the Inversion of Static Vowels

Figure 4.3 shows two examples of static vowel inversion using a linked-codebook on 32 ms, hamming windowed, voiced tokens taken from running speech. The codebook contains 44509 entries pruned from an original 180000 entries and uses the LAM without scaling. The acoustic feature used is the weighted FFT cepstral distance of Meyer et al. [30]. Clearly, linked-codebook lookup alone is often sufficient to get a reasonable acoustic and articulatory fit. Iterative optimization can follow to improve the result if desired.

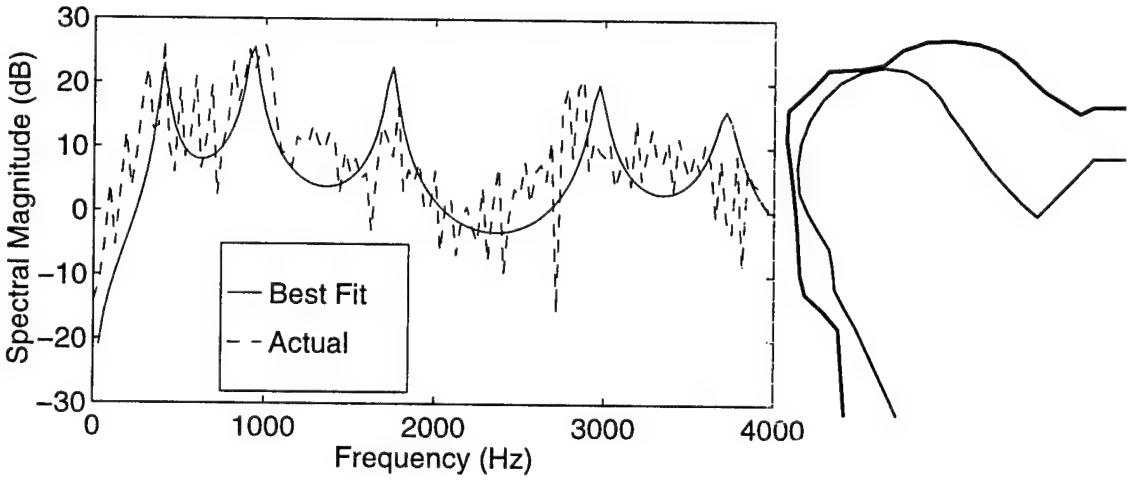
While the match between the synthetic transfer function and the FFT of the signal is close in Figure 4.3(b), the corresponding articulatory configuration has a lip opening for the labial, /w/, at 3.9 cm, that could be larger than desired. If the top three linked-codebook fits are considered, as in Figure 4.4, we find two alternate solutions with smaller lip openings, and similar acoustics. This example illustrates the power of linked-codebooks for acoustic-to-articulatory mapping. The availability of multiple reasonable solutions allows for further processing, especially for dynamic acoustic-to-articulatory mapping, to improve the final result.

4.3.1 Vowel Linked-Codebooks

A great deal of work has been reported on the acoustic-to-articulatory mapping of vowels, so we have not pursued it in depth, preferring to focus our attention on fricatives. We would like to point out a few issues that are not apparent in the



(a) Acoustic and articulatory fit for /i/ of "year".



(b) Acoustic and articulatory fit for /w/ of "were".

Figure 4.3: Results of linked-codebook lookup on two voiced tokens taken from running speech.

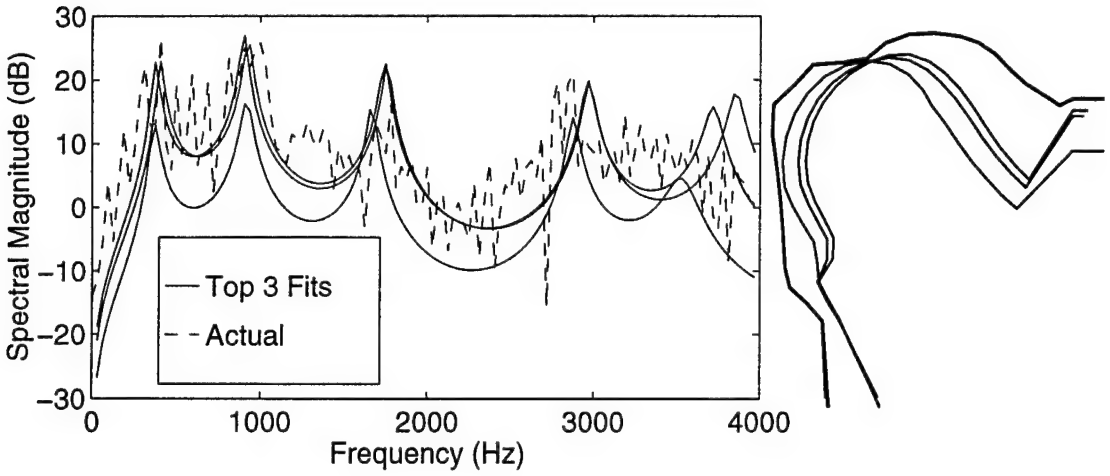


Figure 4.4: Results of linked-codebook lookup: top three acoustic and articulatory fits for /w/ of "were".

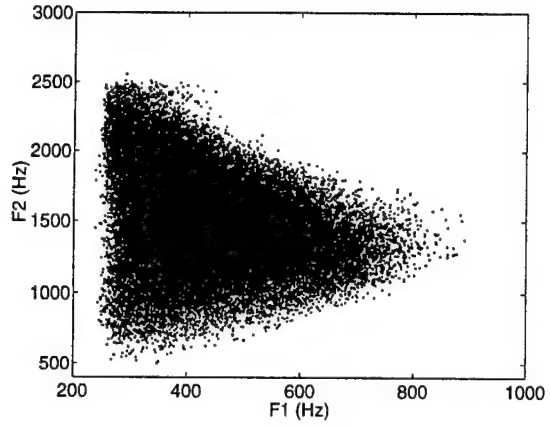
above example, but are unavoidable practical issues in acoustic-to-articulatory mapping for both vowels and fricatives. All are directly or indirectly related to the model mismatch problem.

In the acoustic-to-articulatory mapping of voiced sounds, the vocal-tract shape is typically estimated independently of the glottal-source. This is accomplished by assuming that an appropriately processed speech spectrum represents the vocal-tract transfer function. Separating the source from the vocal-tract in this way is a common approach in speech processing. However, the glottal-source does have an effect on the speech spectrum. Variations in fundamental frequency can bias estimates of the vocal-tract transfer function, especially at low frequencies. Also, the overall spectral slope of the glottal waveform is difficult to separate from the vocal-tract transfer function. Our acoustic-to-articulatory mapping routines were found to be very sensitive

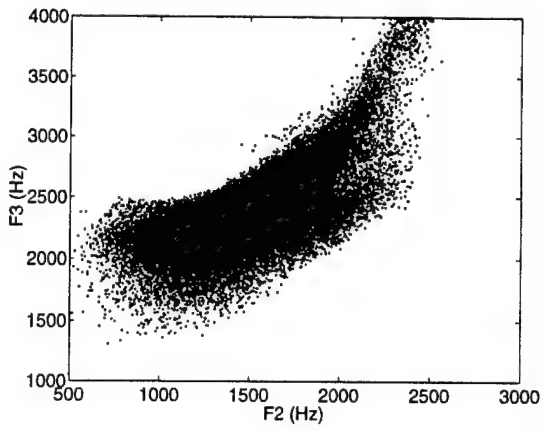
to this spectral "tilt". If the spectral tilt did not match that of the synthesizer's "voice", optimization energy would be spent matching tilt, rather than more perceptually significant aspects. This problem has been observed by other researchers as well [14, 30]. One approach to minimizing the effect of glottal variability is to use acoustic features that are less sensitive to variability in the glottal-source. Weighted cepstral features have been successful in acoustic-to-articulatory mapping and speech recognition for this reason [49].

It is important that the codebook sufficiently cover the acoustic feature space of the test speaker. Figure 4.5 shows scatter plots of formant frequencies calculated for all of the configurations classified as vowel ($A_{vowel} = 0.2cm^2$) compared to formant frequencies measured from samples of speech for 33 adult male talkers in the famous experiment by Peterson and Barney [50]. While the Peterson and Barney dataset does not represent all acoustic possibilities, it does represent features that must be covered by a vowel linked-codebook.

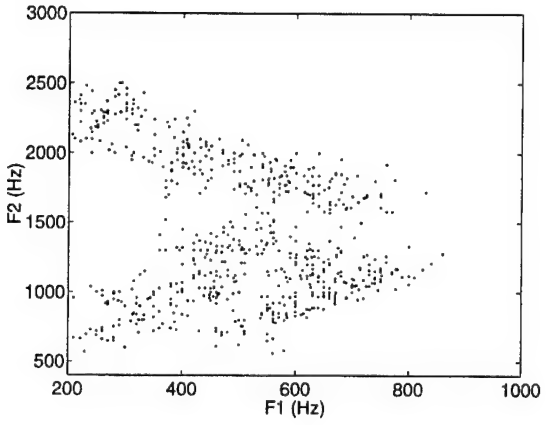
Figure 4.5 suggests that our current vowel codebook is not quite sufficient to cover the formant space. At the left edge of the F1/F2 vowel triangle, F1 does not get below 250 Hz, while the real data have vowels with F1 as low as 200 Hz. This prevents some high vowels from being produced. The upper edge of the codebook F1/F2 triangle does not completely cover that of real speech for large F1 values. Similarly, the lower edge is insufficient for F1 between 500 and 600 Hz. Except for a few points, the remainder of the F1/F2 vowel triangle is sufficiently covered, although the density of samples at the edges is quite low. Another area of concern is the upper edge of the F2/F3 vowel cluster. The limited range of F3 prevents the accurate representation of a large number of vowels. The limited coverage of the formant



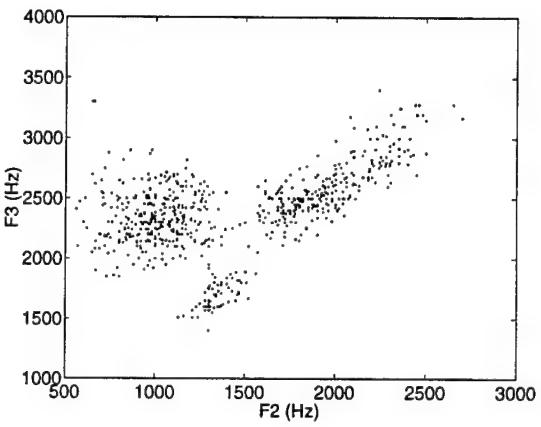
(a) F1/F2, LAM codebook



(b) F2/F3, LAM codebook

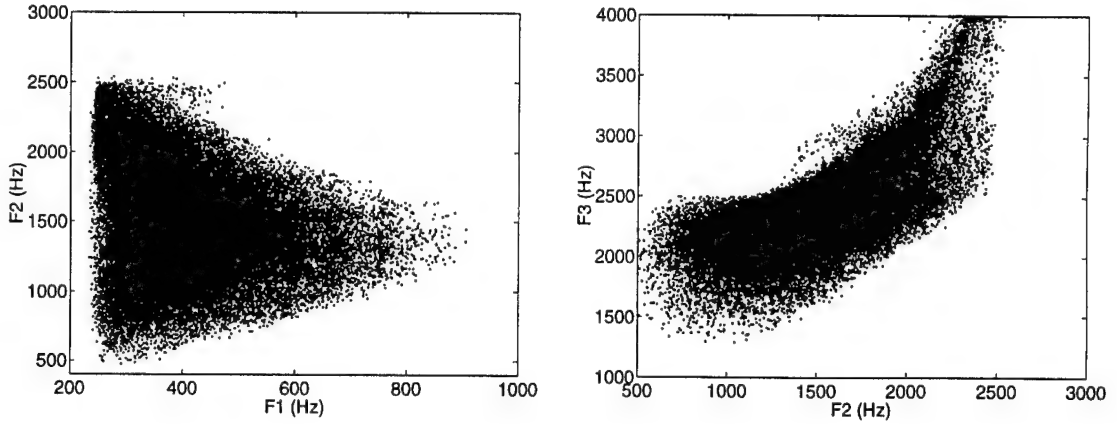


(c) F1/F2, Peterson & Barney



(d) F2/F3, Peterson & Barney

Figure 4.5: Formant frequency scatter plots: comparison of all entries in a vowel codebook using LAM to a collection of spoken vowel samples from Peterson and Barney.



(a) F1/F2 Pruned LAM codebook

(b) F2/F3 Pruned LAM codebook

Figure 4.6: Formant frequency scatter plots after pruning and reducing A_{vowel} to 0.15 for all entries in a vowel codebook using the LAM.

space by the articulatory speech synthesizer is consistent with that reported in the literature for many articulatory models [12, 14, 45]. While the coverage is sufficient for the formant frequencies of prototypical vowels [51], they are not sufficient for all speakers and sounds. Can the codebook's vowel space be expanded and improved in order to be more robust in acoustic-to-articulatory mapping pursuits?

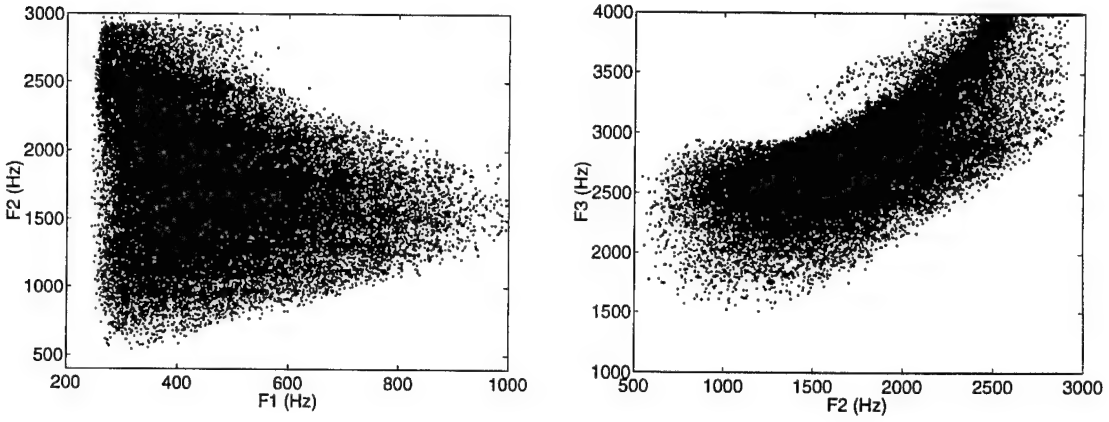
The minimum vowel constriction area threshold, A_{vowel}, has some effect on codebook coverage of the formant space. Reducing A_{vowel} can extend the left edge of the vowel triangle for better coverage of high vowels. To extend the range of F1 down to 200 Hz, we found it necessary to reduce A_{vowel} to 0.05. Unfortunately, a value this small will cause audible frication to be produced for many high vowels. For practical purposes, a compromise of $A_{vowel} = 0.15$ was chosen. As can be seen in Figure 4.6, this choice improves coverage of the high vowels by about 25 Hz without grossly violating the assumptions of our fricative production model. If it becomes necessary,

A_{vowel} can be reduced further and the fricative production model can be adjusted to compensate for vowels with smaller constrictions.

A likely explanation for the discrepancy in the coverage of the formant space between the synthesizer and the Peterson and Barney dataset is the variability between speakers in the length and scale of the vocal-tract. Speakers with shorter vocal-tracts, produce vowels with consistently higher formant frequencies. This was confirmed by including female speakers from the Peterson and Barney dataset into the scatter plots. The vowel triangle expanded significantly, notably to increase the range of F2 and F3. Therefore, with our current articulatory model, there may be certain speakers whose vowels are not adequately covered. The "voice" that is the LAM cannot reproduce the speech of these speakers.

In addition to a restricted coverage of the formant space, preliminary experiments on vowel inversion found that for many speakers, the average spacing between their first four formants is significantly different than the spacing for the synthesizer. This too may be explained by variability between speakers in the length and scale of the vocal-tract. Consider a uniform tube of length l, closed at one end. The resonant wavelengths, $\lambda = (4l, \frac{4}{3}l, \frac{4}{5}l, \ldots)$, are proportional to l. Therefore, the resonant frequencies are inversely proportional to l. Longer vocal-tracts have a smaller average spacing between formants than shorter vocal-tracts.

Although automatically adapting the articulatory model to individual speaker's vocal-tracts is beyond the scope of this work, some speaker normalization techniques were nonetheless required to get reasonable performance in acoustic-to-articulatory mapping experiments. A suboptimal and somewhat inelegant procedure used to get greater coverage of the vowel feature space and better match to real vowel spectra is



(a) F1/F2 Scaled and Pruned LAM codebook

(b) F2/F3 Scaled and Pruned LAM codebook

Figure 4.7: Formant frequency scatter plots after scaling vocal-tract lengths by 0.85 for all entries in pruned vowel codebook using the LAM.

to assume a shorter vocal-tract length in the LAM. This is accomplished by uniformly scaling the length of each section of the area-function by a value slightly less than one. Such an approach is often necessary for any successful inversion to occur; solutions must exist if they are to be found! A more realistic adaptation would include a non-uniform scaling with greater length reduction in the pharynx than in the mouth for female vocal-tracts.

Figure 4.7 shows the vowel triangle after scaling by 0.85. The upper edge of the F1/F2 vowel triangle is extended almost 400 Hz and the right corner is 100 Hz farther to the right. We lose about 50 Hz of coverage on the lower edge of the triangle for small F1. About 25 Hz of coverage is lost on the left edge of the triangle, canceling the gains produced by reducing A_{vowel}. The upper edge of the F2/F3 triangle is extended by almost 500 Hz, with only slight losses on the lower portion of the triangle. This

slight length adjustment to the area-function is sufficient to increase the coverage of the vowel codebook to all of the adult male speakers in the dataset. To cover female speakers adequately, a slightly smaller scale factor is necessary. Some problems may still exist for high vowels with F1 less than 250 Hz.

Our experience suggests that articulatory models with shorter vocal-tract lengths appear to represent the first three formants of long vocal-tracts better than articulatory models with longer vocal-tract lengths represent the formant space of short vocal-tracts. But the average spacing between formants can be too large if the articulatory model's vocal-tract length is too small. Therefore, the best choice of scaling factor appears to be that which produces the longest vocal-tract length that still covers the formant space of the speaker. Selection of a scale factor has been performed by hand when necessary but could be automated.

Not all forms of mismatch between the speaker and the synthesizer can be compensated. Figure 4.8 shows the results of static acoustic-to-articulatory mapping of the onset of /w/ in the word "away" for a male speaker. Of interest in this token is the reduced amplitude of the third formant compared to the second and fourth formants. Our acoustic-to-articulatory mapping procedures are able to match the formant frequencies easily, but are unable to match the formant amplitudes. Further optimization specifically designed to reduce the difference between formant amplitudes, while keeping formant frequencies within 5% of their correct values, was unable to change the formant amplitudes significantly.

Under the right conditions, acoustic-to-articulatory mapping can be achieved in a straightforward manner. The major obstacle and focus of current research is in dealing with undesirable conditions of model mismatch. For voiced sounds, speaker

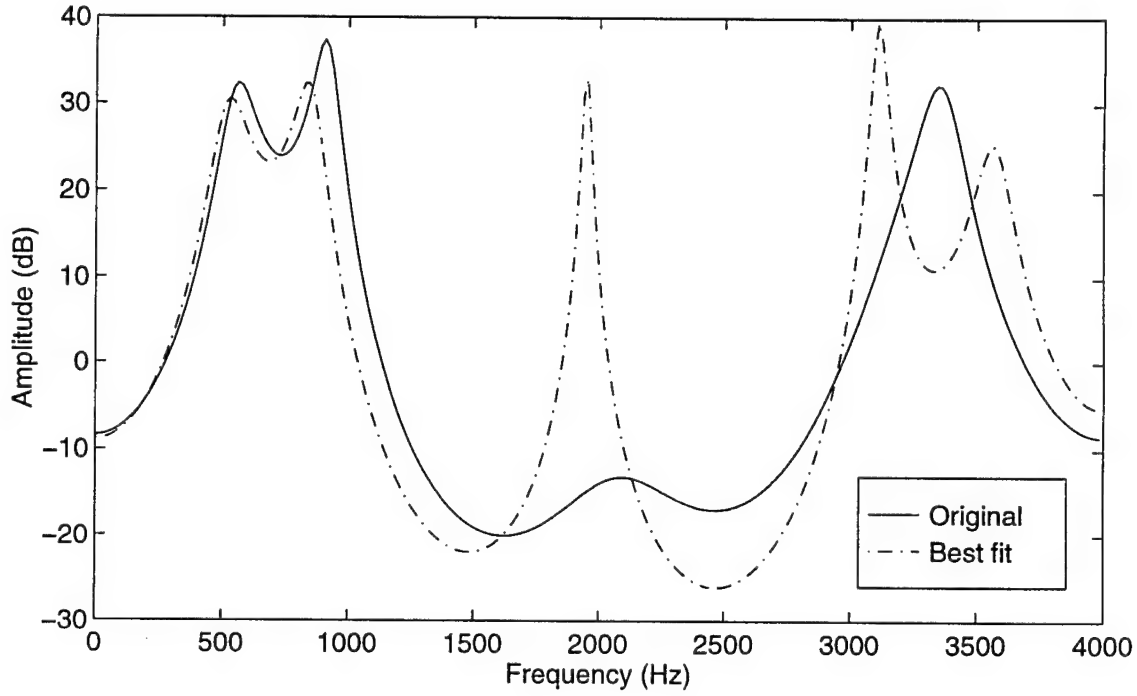


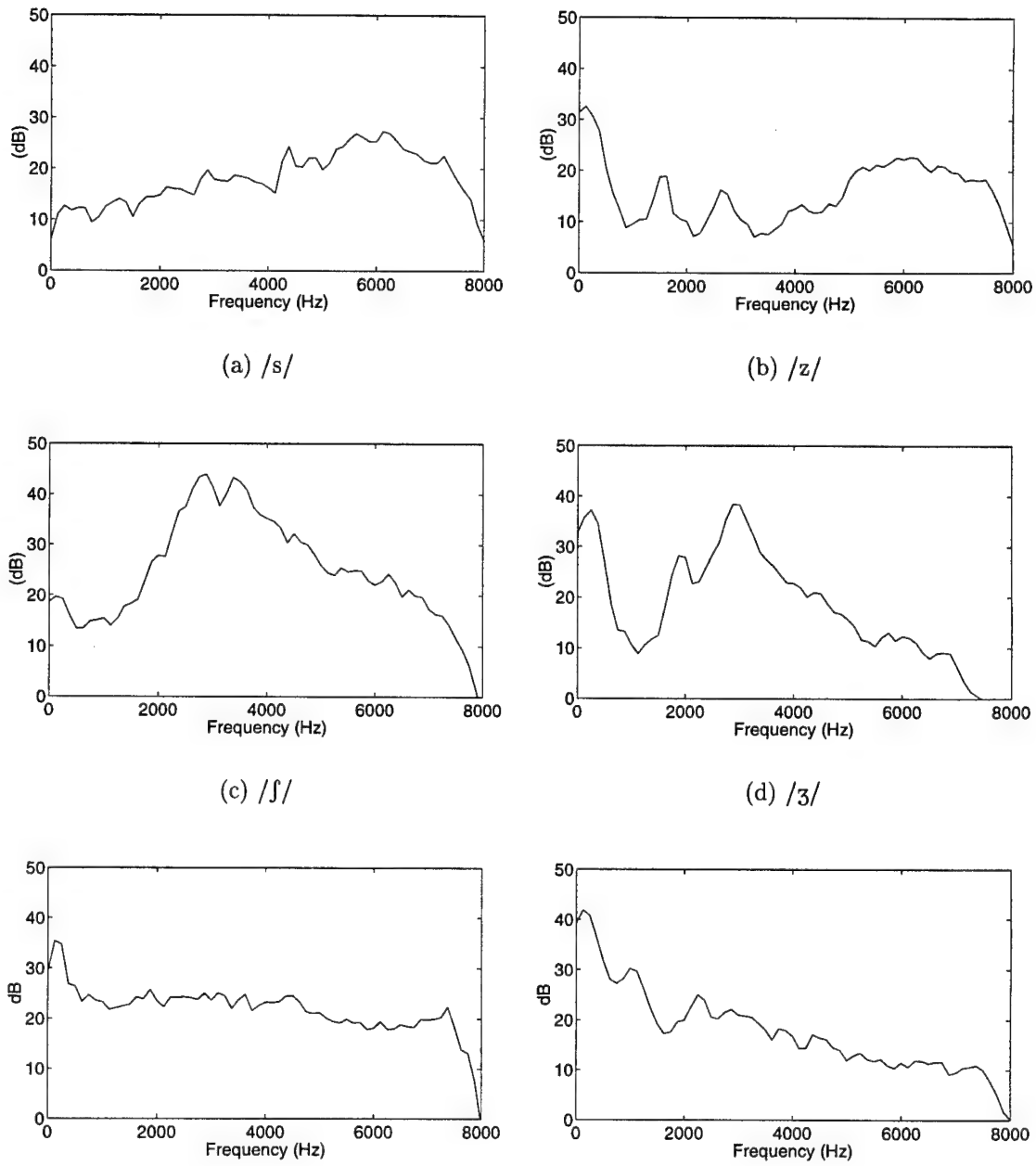
Figure 4.8: Actual and best synthetic fit to LP spectrum of the onset of /w/ in the word "away" for a male speaker. While formant frequencies can be easily matched, the acoustic model is not capable of generating a large difference in formant amplitudes.

adaptation in vocal-tract length and spectral tilt as well as the use of acoustic features resistant to glottal-source variations help reduce the model mismatch problem. For fricatives, the same issues must be addressed, and will have a greater influence on the success of acoustic-to-articulatory mapping due to the greater uncertainty in the fricative production model

4.4 Linked-Codebook Lookup for the Inversion of Static Fricatives

While extension of the linked-codebook procedure to fricatives may seem straightforward, there are a number of issues that make the task more challenging. The selection of effective acoustic features and efficient linked-codebook sampling and pruning techniques for fricatives requires knowledge about the articulation, acoustics, and perception of fricatives. Increased amounts of model mismatch and source-tract interaction, as well as the mixing of phonation and frication require creative solutions.

Fricatives are distinguished by the presence or absence of voicing and their place of articulation. Spectrally, Strevens [52] described the spectral shape of fricatives by separating fricatives into three groups: front, middle, and back fricatives. Front fricatives, /f,v,θ,ð/, have the lowest intensity and smoothest spectra. Middle fricatives, /s,z,ʃ,ʒ/, have a highest intensity and have spectra characterized by one or more major "humps" in the middle frequency range. Although not considered here, back fricatives such as /x/ and /χ/ have moderate intensity and spectra with a formant-like structure. Figures 4.9 and 4.10 illustrate the spectra of some English fricatives for a male and a female speaker. The spectra of voiced fricatives contain energy at low frequencies due to the modulation of airflow by vocal-fold vibration. At higher frequencies, above the fourth or fifth harmonic of the fundamental frequency, voiced and

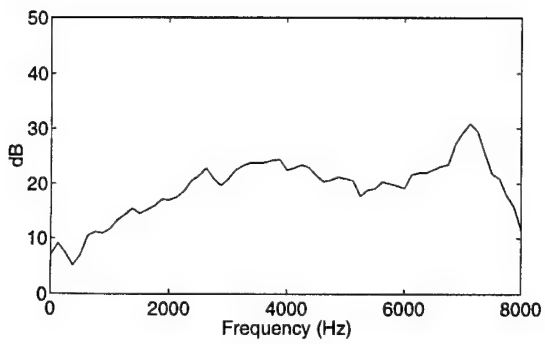


(a) /s/ (b) /z/

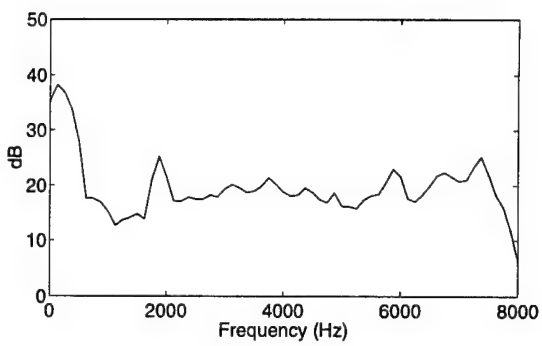
(c) /ʃ/ (d) /ʒ/

(e) /f/ (f) /v/

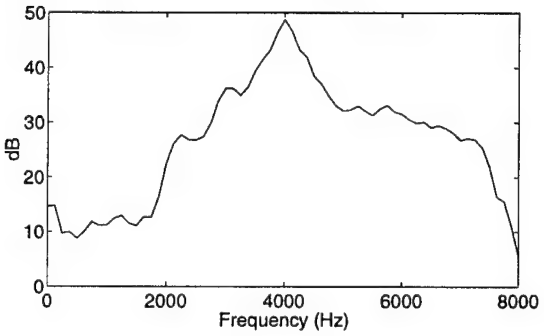
Figure 4.9: Fricative spectra for speaker MJC.



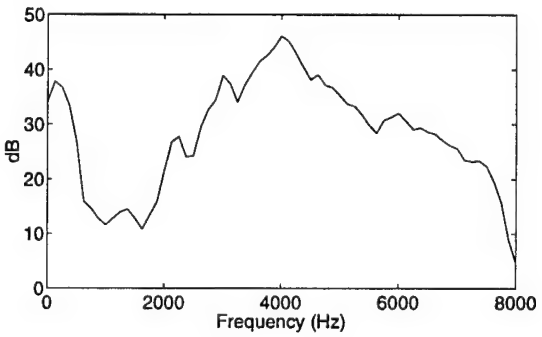
(a) /s/



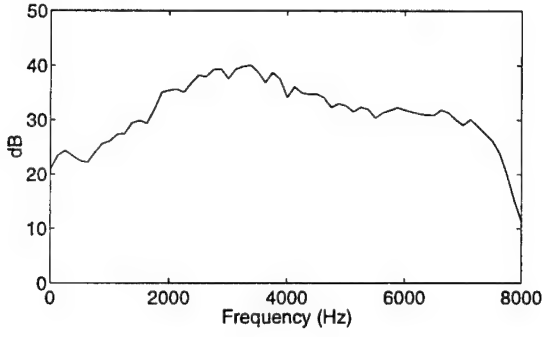
(b) /z/



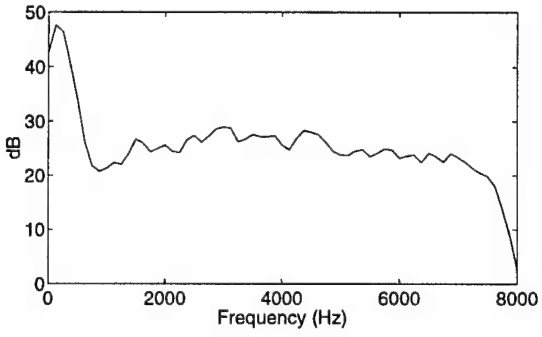
(c) /ʃ/



(d) /ʒ/



(e) /f/



(f) /v/

Figure 4.10: Fricative spectra for speaker WLR.

unvoiced fricatives of the same place of articulation have quite similar spectra [51]. Hughes and Halle [53] found that spectral distinctions between fricatives are more consistent within a given talker than across talkers. Spectral variation in fricatives across talkers can be so great that Hughes and Halle report, *"The discrepancies among the spectra of a given fricative as spoken by different speakers in different contexts are so great as to make the procedure of plotting these spectra on one set of axes a not very illuminating one. On the other hand, the differences among the three classes of fricatives (labial, dental, and palatal) are quite consistent, particularly for sounds spoken by a single speaker."*

Perceptual experiments have found that fricatives are not as well distinguished as vowels by listeners, and that listeners rely on cues in addition to the fricative spectra, such as relative amplitude and formant transitions into and out of the fricative to make their decisions [54, 55]. An experiment by Harris [55] supports this conclusion. Listeners were asked to classify fricative-vowel tokens in which the noise from the fricative portion was substituted with noise from other tokens. The results suggest that listeners use spectral information to distinguish the fricatives /f,s,ʃ/, but require additional cues from the formant transitions to distinguish between the front fricatives /f/ and /ð/. Similar results were found for voiced fricatives.

For vowels, the sole acoustic energy source is at the glottis. For voiced fricatives, acoustic energy from frication is mixed with glottal energy. Even unvoiced fricatives contain some voicing, especially intervocalic unvoiced fricatives. Transitions from unvoiced fricatives to vowels require a duration of mixed frication and phonation. Resolving this mixture is a serious challenge for the acoustic-to-articulatory mapping of fricatives.

While much about the production of vowels is known, much less is known about the production of fricatives. While the basic process is understood [47, 23], accurate models of turbulent flow and the noise generated by that flow are not available. Linear acoustic wave propagation, an assumption in most articulatory speech synthesizers, breaks down near the constriction. It is believed that the noise source spectrum may not be the same for all fricatives, and may vary as a function of airflow. There may be multiple noise sources, or distributed sources. Recent research has attempted to identify the spectral properties of the fricative noise source [56, 57, 58, 59, 48], but much work remains to be done.

The development of a better model of frication is beyond the scope of this work. Except for minor adjustments, we intend to solve the inverse problem for the forward model as defined in Chapter 3. Some acoustic features specifically designed for fricatives will be investigated. Our experiments will verify and extend the work on the acoustic-to-articulatory mapping of unvoiced fricatives by Sorokin [17] and Shirai [18]. Some of the more difficult issues of mixed modes and source-tract interaction will be addressed in the dynamic fricative inversion studies of Chapter 6.

4.4.1 Fricative Linked-Codebooks

The generation of fricative linked-codebooks follows the same procedures as for vowel linked-codebooks, but with sampling, pruning, classification, and feature extraction tailored to fricative production, acoustics and perception. A wealth of knowledge and experience in the processing of fricated speech does not exist like it does for voiced speech. Therefore, the generation of good fricative codebooks will require more experimentation.

The linear articulator model used in the vowel case, will also be used as an articulatory representation for fricatives. One additional parameter, fricative noise source location, is added to give the model eight parameters in all. The fricative noise source location is defined relative to the constriction. This modified model is referred to as the fricative linear articulator model (FLAM).

Articulatory codebooks were generated using the FLAM representation without any additional source parameters. While significant source-tract interaction exists for fricatives, we believe that linked-codebooks, designed without regard to the relation between constriction area and frication amplitude, will still provide good starting points for further optimization. Uniformly distributed random sampling was used to collect 20000 fricative configurations, with $A_{fric} = 0.3cm$. Codebook pruning was not performed due to a lack of a justifiable pruning criteria.

Effective acoustic features for fricatives have not been established since fricatives are seldom analyzed individually in speech processing applications. Based on our knowledge of fricative acoustics and perception, a good fricative acoustic feature should represent the overall spectral shape and distribution of energy; however, high resolution in frequency does not appear to be necessary. The acoustic distance measures used by Shirai and Masaki [18] and Sorokin [17] in their respective static fricative acoustic-to-articulatory mapping experiments satisfy these criteria. Therefore, the same types of features, described below, are used in our linked-codebooks.

1. FWCEP: This is a version of the weighted FFT-cepstral feature of [30]. Since our fricatives are sampled at 16 kHz, the resampled weighting function

$$w_i = \begin{cases} (i/40)^{0.4}, & 0 < i \leq 40 \\ 0.5 + 0.5\cos(\pi(i-41)/40), & 40 < i \leq 60 \end{cases} \tag{4.5}$$

is double the original feature length. During linked-codebook lookup, a Euclidean distance is used with this feature.

2. FPSD: This is a 64 point power spectral density sampled linearly over the 0–8 kHz range, excluding the DC term. During linked-codebook lookup, normalized correlation,

$$Q = \frac{\int_0^{\omega_s/2} S_1(\omega) S_2(\omega) d\omega}{[\int_0^{\omega_s/2} S_1(\omega) d\omega \int_0^{\omega_s/2} S_2(\omega) d\omega]^{1/2}}, \tag{4.6}$$

is maximized, where S_1 is the synthetic spectrum and S_2 is the actual spectrum.

The spectra of voiced fricatives differ primarily from that of unvoiced fricatives by the presence of energy at low frequencies due to the modulation of airflow by vocal-fold vibration. At higher frequencies, above the fourth or fifth harmonic of the fundamental frequency, voiced and unvoiced fricatives of the same place of articulation have quite similar spectra [51]. In an attempt to produce acoustic features that are somewhat insensitive to the presence of voicing, the above two acoustic features were modified. The new FWCEP_V and FPSD_V features are identical to the FWCEP and FPSD features respectively, except that the acoustic feature is calculated only over the 1–7.5 kHz range.

4.4.2 Analysis of Fricative Linked-Codebooks

In the classical definition of acoustic-to-articulatory mapping, the distance between the estimated and actual area-functions is minimized. Unfortunately, this metric is difficult to employ on natural speech, since the true area-function is unavailable. For investigative purposes, an articulatory speech synthesizer can be used to provide speech with known area-functions for acoustic-to-articulatory mapping experiments. This allows for performance evaluation isolated from the negative effects

of model mismatch and measurement noise. Evaluation using synthetic speech can be interpreted as an upper bound on performance, since model mismatch and noise will degrade results.

A measure of how closely an inverse mapping procedure matches the true configuration can be obtained by using a sample of synthetic speech whose generating configuration is known. But the conclusions we can draw from a single instance are few. By repeating the experiment for a large number of samples, consistent and insightful statistics can be generated. Such a technique is used to examine fricative linked-codebooks.

Given an articulatory vector, $\mathbf{p}$, and its acoustic feature vector consequence, the *articulatory error*, $\mathbf{e}$, between an acoustic-to-articulatory mapping result, $\mathbf{p}_{n^*}$, and the actual articulatory vector is

$$\mathbf{e} = \mathbf{p} - \mathbf{p}_{n^*}. \tag{4.7}$$

Let *articulatory distance*, d, be defined as the Euclidean distance between the result and the actual articulatory vector.

$$d = \sqrt{\sum_{i=1}^{N} e_i^2} \tag{4.8}$$

If we assume that the input articulatory vector is a random variable, uniformly distributed over the range of reasonable values in the articulatory space, then articulatory error and articulatory distance are random variables.

The average value of articulatory distance, $d_{avg} = E\{d\}$, as well as its one dimensional probability density function, illustrates the performance of the acoustic-to-articulatory mapping routine and allows comparisons among different algorithms and

features. The covariance of the articulatory error,

$$\mathbf{K} = E\{(\mathbf{e} - E\{\mathbf{e}\})(\mathbf{e} - E\{\mathbf{e}\})^T\} \tag{4.9}$$

provides information about how the error is distributed, and helps to identify which dimensions contribute most or least to the error.

Acoustic Feature Evaluation

The ability of linked-codebooks to represent the acoustic-to-articulatory transformation was evaluated for different acoustic features by examining the distribution of articulatory distance for a large number of samples. For each of the four linked-codebooks described previously, articulatory error vectors were collected from the linked-codebook lookup of 20000 randomly generated fricative configurations. The random configurations were distributed uniformly over the same range of input parameters used to generate the codebooks. Additionally, articulatory error and distance were collected for "best" and "random" lookup. Best lookup chooses the articulatory configuration in the linked-codebook that is closest to the true configuration. Although not ideal performance, best lookup performance is a measure of the quantization noise inherent in the linked-codebook representation. Random lookup randomly chooses an articulatory configuration in the linked-codebook, assuming every codeword to be equally probable. For the calculation of articulatory error and distance, each articulatory parameter was normalized to the range of 0 to 255.

The distributions of articulatory distance for all four acoustic lookups, along with best and random lookup, are displayed in Figure 4.11. The curves are normalized histograms of the data, and approximate the probability density function of the articulatory distance for the codebook. Distributions located farther to the left indicate

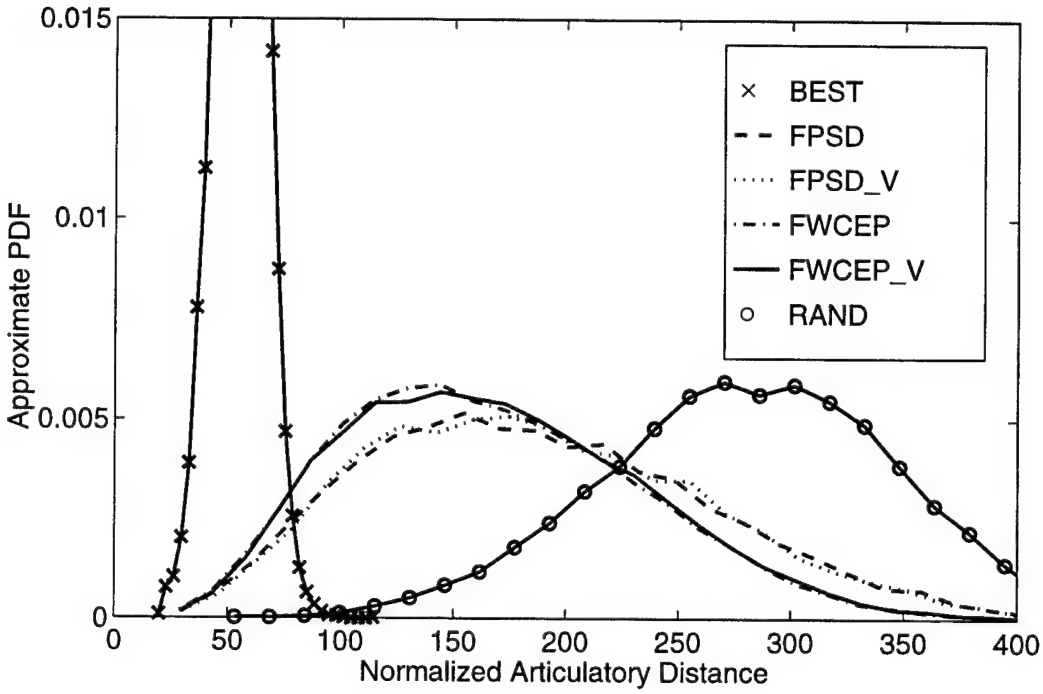


Figure 4.11: Normalized histograms of articulatory distance for all acoustic lookups compared to histograms for best lookup and random lookup.

better acoustic-to-articulatory mapping performance. All acoustic lookups outperform random lookup but remain distant from best lookup. Acoustic lookups with the FWCEP feature appear to be better than acoustic lookups with the FPSD feature. This is true whether Euclidean distance or normalized correlation is used. For both feature classes, the loss of information from below 1 kHz in the "voiced" versions has little effect on the distribution of articulatory distance.

Error in best lookup is due to quantization of the articulatory space by the linked-codebook. Acoustic lookup also leads to error in the articulatory domain, but this error comes about because of quantization of the acoustic space. The articulatory error for acoustic lookup will, in general, not be distributed uniformly as for best lookup. On average, the articulatory error for acoustic lookup will vary across dimensions due to the relative sensitivity of the acoustic representation to different articulatory dimensions. The following section investigates this distribution of articulatory error in acoustic lookup and identifies articulatory dimensions of fricatives to which the acoustic representation is most sensitive.

Investigation of Error Variance

The distribution of articulatory error for the four linked-codebooks was investigated by examining the articulatory error covariance matrix in each case. The covariance matrices for best and random lookup resemble scaled identity matrices with averaged scaling factors of 387 and 10430 respectively. Since all of the articulatory dimensions have been normalized to the same range, the matrices suggest that estimation error in the best and random cases is distributed evenly between the articulatory dimensions. The covariance matrices for the acoustic lookups indicate that the error is not distributed evenly among the articulatory dimensions. Comparing the

Lookup Type	Jaw Angle (JAW)	Tongue Body Position (TBP)	Tongue Body Shape (TBS)	Tongue Tip (TT)	Lip Height (LPH)	Lip Prot. (LPP)	Larynx Height (LRH)	Fric. Loc. (FSL)
BEST	3.97	3.57	2.78	3.82	4.63	3.83	3.83	3.85
FPSD	66.53	41.41	23.90	64.55	19.24	50.10	71.20	54.10
FWCEP	56.25	24.56	12.12	54.43	9.07	45.96	68.06	35.35
FPSD_V	65.62	39.52	22.93	63.83	17.11	50.67	74.39	54.10
FWCEP_V	56.69	25.13	12.87	54.47	8.46	45.92	73.31	34.85
RAND	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Table 4.2: Error variance in FLAM dimensions for all lookups, normalized by random lookup variance.

estimation error in different articulatory dimensions is difficult using the covariance matrices alone, since the range over which each parameter may be varied has some influence on the error. On the other hand, comparing the articulatory error in each dimension to the corresponding values for best and random lookup is a more reliable and fair method of comparison.

Table 4.2 shows the variance of the articulatory error in different articulatory dimensions for all four acoustic lookups compared to best and random lookup. The variances are normalized by the variance for random lookup. Notice that the variances in all dimensions are smaller for the FWCEP features than the FPSD features, echoing our histogram observations. For all acoustic lookups, the variances of LPH, TBS, and TBP are small compared to the large variances of LRH, JAW, and TT. Apparently, the linked-codebook lookup can estimate certain dimensions more accurately than others. An alternate, equivalent explanation is that some articulatory dimensions have a greater effect on the acoustics than others.

Lookup Type	Constriction Location	Frication Location	Vocal-Tract Length	Vocal-Tract Area Above Glottis
BEST	14.12	11.25	3.95	4.97
FPSD	12.99	8.21	46.28	80.48
FWCEP	1.55	4.61	33.95	67.21
FPSD_V	13.14	9.90	47.79	81.62
FWCEP_V	2.26	6.64	36.36	76.11
RAND	100.0	100.0	100.0	100.0

Table 4.3: Error variance in different articulatory dimensions for all lookups, normalized by random lookup variance.

The disparity in estimation ability is much more obvious when measured in articulatory dimensions significant to fricatives. Constriction and frication location, which define the front and back cavities for fricative configurations, are known to have a great influence on the fricative transfer function. As shown in Table 4.3, these dimensions have, for all acoustic lookups, very low normalized variances. In fact, they are smaller than the variance *for best lookup*. This is possible because best lookup minimizes articulatory error across all FLAM dimensions. The codeword that minimizes the distance in one dimension will not, in general, be the same codeword that minimizes distance across all dimensions.

Also shown in Table 4.3 are the normalized variances for vocal-tract length and vocal-tract area directly above the glottis. The variance is quite large for these cases. Since these dimensions only affect the back cavity, they have less influence on the acoustics and as a result, are less well estimated.

In Table 4.3, the variance for best lookup across articulatory dimensions is larger for constriction location and frication location (14.12 and 11.25 respectively) as compared to the values for vocal-tract length and area above the glottis (3.95 and 4.97 respectively). As described in Section A, constraints were placed on constriction area, constriction location, and frication location during sampling of the articulatory space so that only valid fricative configurations were retained. These constraints restrict the range of parameter choices, without affecting the coarseness of the codebook quantization. Therefore, the variance for random lookup and acoustic lookup is reduced, while variance for best lookup is unchanged. In order to make fair comparisons between articulatory dimensions that may have been constrained differently, lookup variances have been normalized by random lookup variance. As a result, normalized variance for best lookup is larger for dimensions that have been constrained.

The relative accuracy of articulatory estimates can be employed in any aspect of acoustic-to-articulatory mapping that considers articulatory distance. Articulatory distance is used as a form of regularization in static inverse mapping [24] and as a measure of continuity in dynamic inverse mapping [25]. The ability to weight distance or enforce continuity more strongly in the dimensions whose estimates have the greatest confidence will help to focus optimization energy and improve results.

4.4.3 Linked-Codebook Performance on Real Fricatives

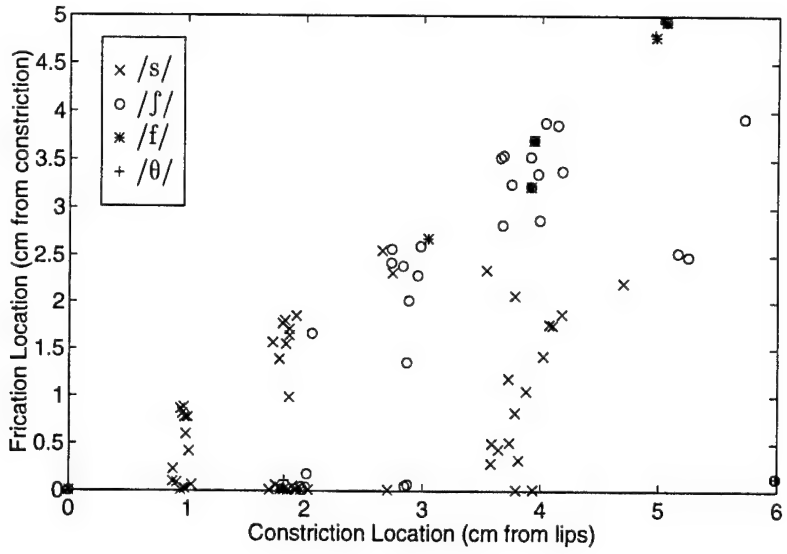
A corpus of fricative data has been collected consisting of VFV (vowel-fricative-vowel) clusters and fricatives in isolation. The point vowels (/i/, /u/, /a/) were used for VFV clusters of large contextual contrasts. English fricatives were collected from four native English speakers, two male and two female. Specifically, the corpus

consisted of the unvoiced fricatives /s/, /ʃ/, /f/, /θ/, and the voiced fricatives /z/, /ʒ/, /v/, /ð/. The utterances were recorded in a sound booth using a head mounted microphone. Two repetitions of each sample, in both voiced and whispered contexts, were taken for a total of 384 fricatives recorded. Fricatives in the corpus were sampled at 16 kHz and segmented into 64 ms frames with a 32 ms overlap for a total of 2217 tokens.

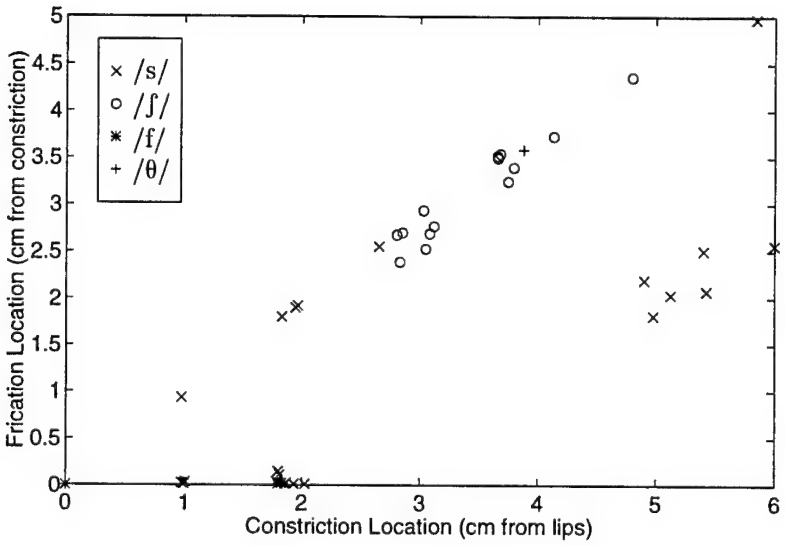
Acoustic-to-articulatory mapping using linked-codebooks is examined for unvoiced and voiced fricatives from this corpus. If inverse mapping is working correctly, the resulting articulatory configurations of a speaker should be consistent within phonetic class, i.e., fricatives of the same phonetic class should be produced by similar articulatory configurations, while configurations for different phonetic classes should be significantly distinct. Since fricatives are best distinguished by their place of articulation, we expect the lookup results to be the most consistent in that dimension. Therefore, scatter plots of acoustic-to-articulatory mapping results in constriction location, along with frication location, will be examined.

Unvoiced Fricatives

Linked-codebook lookup using the four linked-codebooks was performed for all of the unvoiced tokens of one male speaker from our corpus. Lookup was allowed to select the best codebook fit over a rather wide range of constriction and frication locations. Figure 4.12 shows scatter plots for two of the codebooks. Note the vertical striation in both plots. This is a consequence of the FLAM, which has fixed length (1 cm) sections in the area-functions it produces. In Figure 4.12(a), we see that the FPSD lookup produces reasonable separation by phonetic class, but with multiple clusters for some classes. This is not consistent with natural fricative production. While there will be



(a) FPSD lookup



(b) FWCEP_V lookup

Figure 4.12: Clustering in constriction and frication location of linked-codebook lookup results on unvoiced fricative tokens of speaker MJC.

some variation in articulation within a phonetic class due to coarticulation, we do not expect distinctly different forms of articulation. FPSD lookup also has a tendency to form configurations with frication locations near the lips, which is physically unlikely when the constriction location is large. Generally, the acoustics of the results do resemble the real fricatives they model, although results with large frication locations exhibit acoustic fits with a large dynamic range and deep troughs at low frequencies. FWCEP lookup did not perform well, employing only a limited number of codewords with frication locations less than 0.1 cm. FWCEP_V lookup performs much better and is shown in Figure 4.12(b). Like FPSD lookup, FWCEP_V lookup shows good class separation, although, again, with multiple clusters for some classes. All configurations for /ʃ/ have frication locations at the lips, which is physically unlikely.

For both linked-codebook lookups, the majority of dental and labio-dental fricatives, /θ/ and /f/, are located at the lips and have reasonable acoustic fits. While these fricatives are not produced precisely at the lips, this may be a good choice for FLAM. The constriction locations for /θ/ and /f/ should be very similar and located forward of constrictions for /s/. The articulatory characteristic that distinguishes /θ/ from /f/ in a one-dimensional acoustic model is tongue position, and the linked-codebook cannot consistently make this distinction. Perceptually, these two sounds are hard to distinguish, so it is no surprise that the acoustic-to-articulatory mapping algorithm cannot separate them. Contextual cues may be necessary to correctly separate the two.

Unvoiced Fricatives with Constraints

The analysis of frication and constriction location shows that our scattering results suffer from extreme values, and do not exhibit the unimodal clustering we expect.

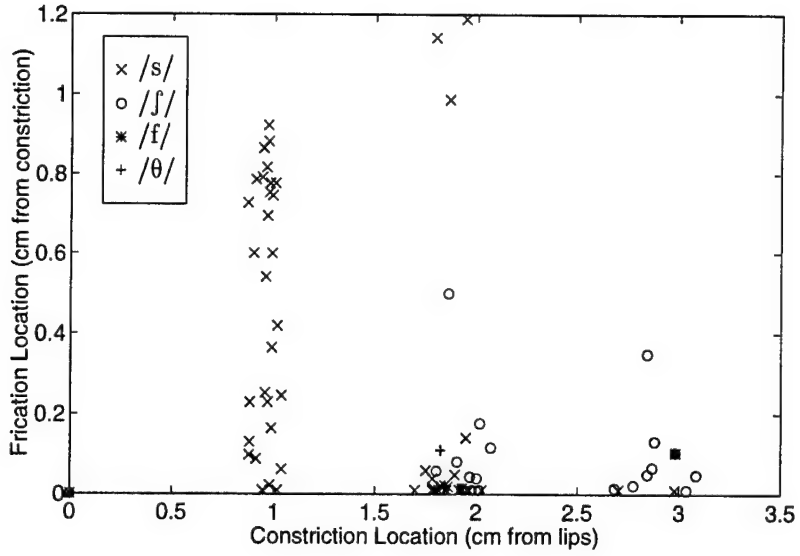
Therefore, constriction and frication location have been constrained. Constriction location is limited to 3.5 cm, just large enough to cover the English fricatives in our dataset. Frication location is limited to 1.2 cm in front of the constriction. These constraints reduce codebook size by about 40%.

Figure 4.13 shows scatter plots of codebook lookup with these constraints on constriction location and frication location. With the additional constraints we see a significant improvement in the clustering results. The /s/ and /ʃ/ clusters, along with the /f/-/θ/ cluster at the lips, are distinguished by constriction location. There is some overlap between the clusters for /ʃ/ and /s/ that may be due the inability of the FLAM to define constriction location continuously.

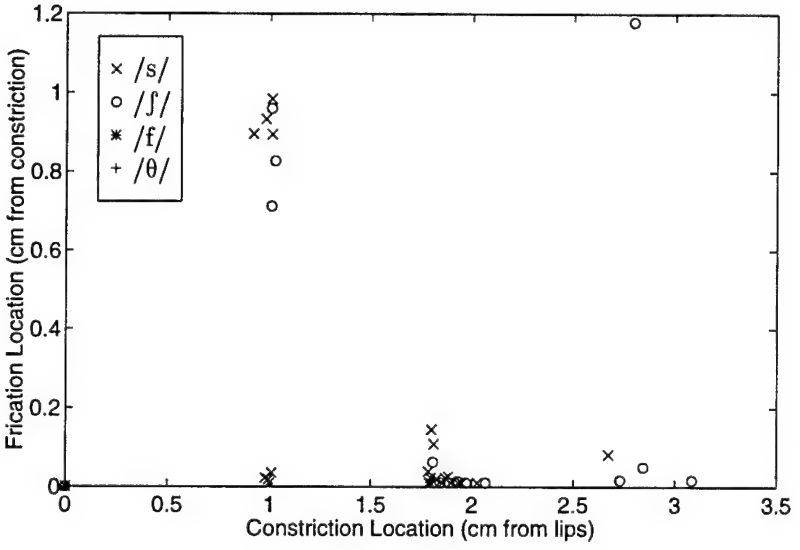
Figures 4.14 and 4.15 show examples of FPSD lookup. The acoustic fitness for /s/, /f/, and /θ/ in Figures 4.14(a), 4.15(a), and 4.15(b) is good. The synthetic fricative spectrum for /ʃ/ in Figure 4.14(b) has a formant-like structure, unlike the real fricative spectrum. This may be due to the inability of the FLAM to produce a cavity under the tongue when the tongue is raised to the palato-alveolar position. Fricatives synthesized from these results sound reasonable, even for /ʃ/, but sound quality is difficult to judge in isolation. Results using FWCEP_V lookup are comparable.

Voiced Fricatives

Figure 4.16 contains scatter plots of results on voiced fricatives for the same male speaker using FPSD and FPSD_V lookup. FPSD lookup performs worse for voiced fricatives than for unvoiced fricatives. Apparently, the presence of low frequency voicing energy in the spectrum prevents FPSD lookup from distinguishing fricatives. In general, fewer codewords are used for the voiced fricatives. Only 5 codewords not located at the lips were used for voiced alveolars. Furthermore, configurations were

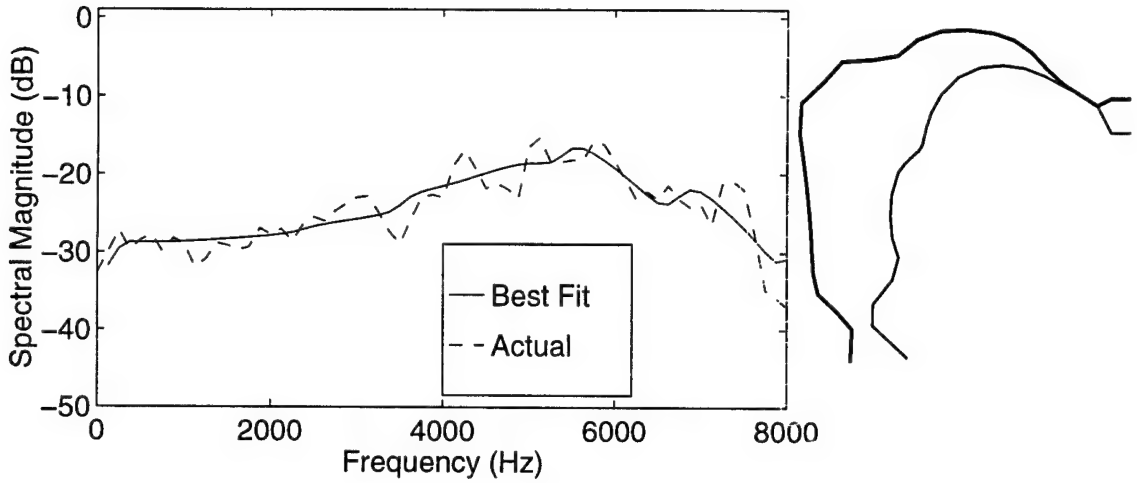


(a) FPSD lookup

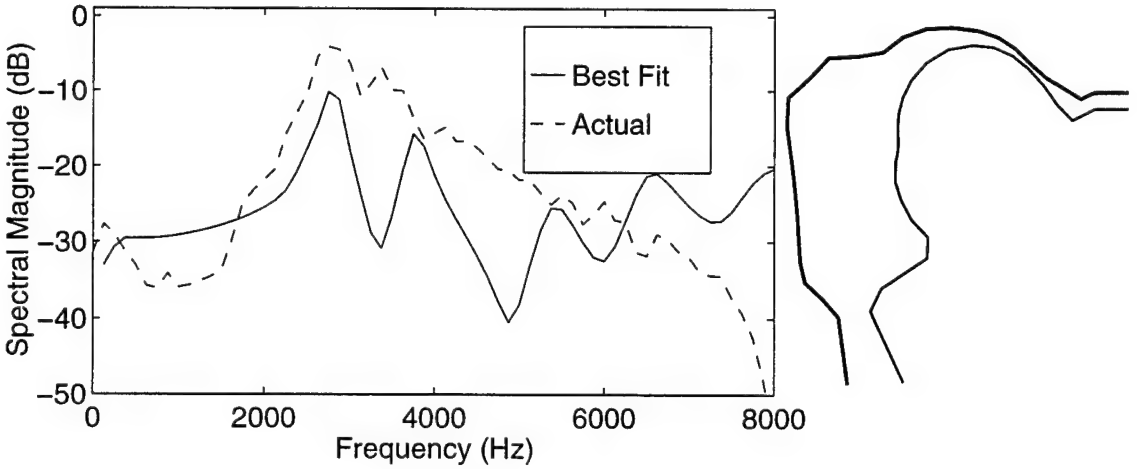


(b) FWCEP_V lookup

Figure 4.13: Constrained clustering in constriction and frication location of linked-codebook lookup results on unvoiced fricative tokens of speaker MJC.

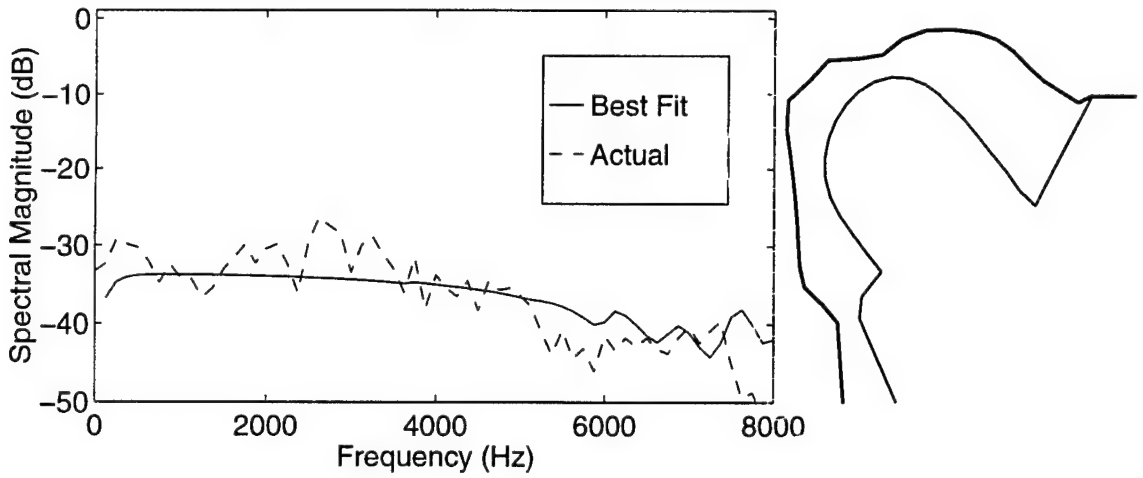


(a) /s/

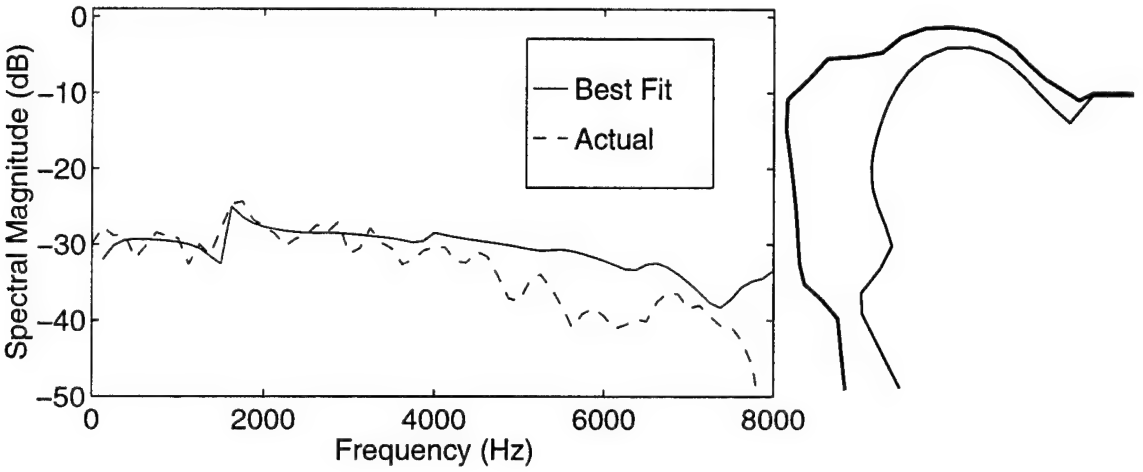


(b) /ʃ/

Figure 4.14: Acoustic fit and articulatory configuration for /s/ and /ʃ/ using FPSD lookup.

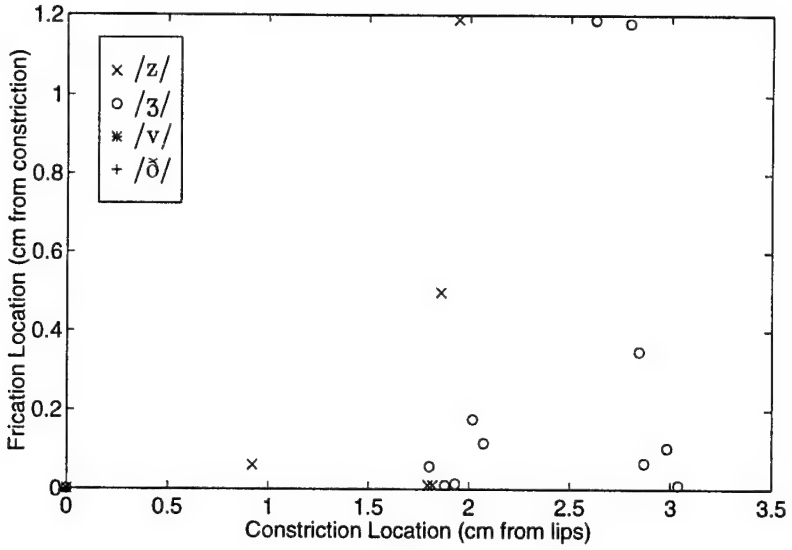


(a) /f/

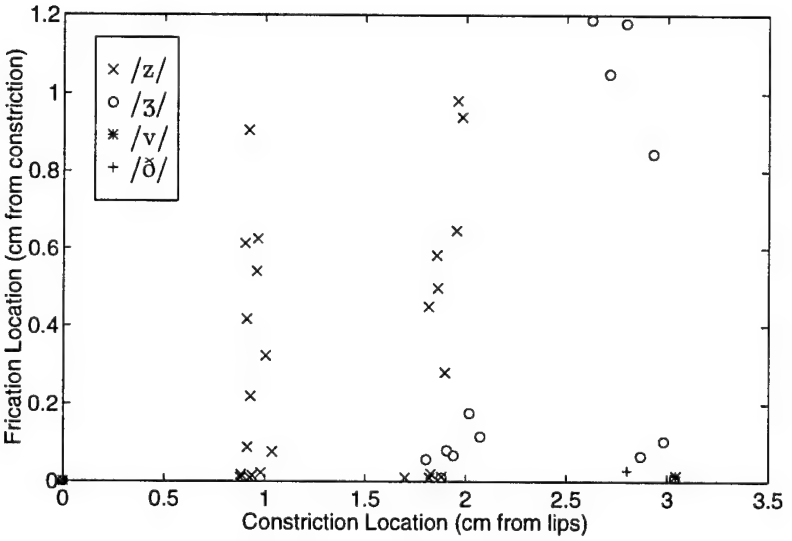


(b) /θ/

Figure 4.15: Acoustic fit and articulatory configuration for /f/ and /θ/ using FPSD lookup.



(a) FPSD lookup



(b) FPSD_V lookup

Figure 4.16: Constrained clustering in constriction and frication location of linked-codebook lookup results on voiced fricative tokens of speaker MJC.

located at the lips for 86% of the voiced alveolar fricatives. Codeword usage for the voiced palato-alveolars are similarly reduced, although not as many palato-alveolars are located at the lips. Clustering for FPSD_V lookup in Figure 4.16(b) is much improved, and resembles the clustering performance and acoustic fitness of FPSD lookup for unvoiced fricatives. By eliminating contributions from frequencies below 1 kHz, the FPSD_V feature appears to better represent voiced fricatives. Linked-codebook lookup results for voiced fricatives using FPSD_V lookup are comparable to the unvoiced examples given in Figures 4.14 and 4.15.

4.5 Discussion

We believe that the quality of our inverse mapping for fricatives is sufficient to be of value to a dynamic acoustic-to-articulatory mapping system. The ability of linked-codebooks to provide multiple valid solutions combined with continuity constraints can overcome the occasional poor choice by the linked-codebook. Chapter 5 discusses the application of linked-codebooks to the dynamic acoustic-to-articulatory mapping of voiced sounds. The dynamic acoustic-to-articulatory mapping procedures are then extended to include fricated speech in Chapter 6 using the features and constraints developed in this chapter.

CHAPTER 5

AN ACOUSTIC-TO-ARTICULATORY MAPPING SYSTEM FOR VOICED SOUNDS

5.1 Introduction

In the previous chapters, we have considered only the static case of acoustic-to--articulatory mapping. This chapter considers the dynamic acoustic-to-articulatory mapping problem for voiced utterances. As mentioned in Chapter 4, voiced speech refers to only vowels and glides, without liquids or any fricated, aspirated, nasalized, or plosive sounds. Acoustic-to-articulatory mapping for voiced sounds is a logical first step in the development of a complete inverse mapping system for all speech sounds. Dynamic inversion for voiced sounds has been attempted by a number of researchers with some success (see Chapter 2 for a review of past work). Many of the procedures described in herein are based on this previous work.

A frame-based approach has been taken to the dynamic mapping problem. In this approach, the continuous articulatory trajectory, $\mathbf{p}(t)$, is represented with a uniformly sampled version, $\mathbf{p}_m$, such that

$$\mathbf{p}_m = \mathbf{p}(t)|_{t=mT}, \quad m = 1, \ldots, M, \tag{5.1}$$

where T is the sampling period and M is the total number of frames. Each sample represents a frame over which a single articulatory configuration is estimated, much

like in the static case. Sampling periods are generally in the range of 5 ms to 30 ms. Since our articulatory speech synthesizer operates in the frequency-domain and is frame-based, the frame-based approach is logical and straightforwardly implemented.

As in Chapter 4, acoustic-to-articulatory mapping on each frame will be accomplished using linked-codebook lookup. The linked-codebook lookup procedure can be used to find the N codebook configurations that most closely resemble a given speech segment according to some distance metric. The results of this N-best codebook lookup are candidate articulatory configuration solutions and may be distributed across the entire articulatory space due to the one-to-many nature of the acoustic-to-articulatory transformation. Instead of merely selecting the best linked-codebook fit as in the static case, continuity constraints can be applied to select from the N-best candidates in each frame the best trajectory matching the given utterance. Constraints on frame-to-frame transitions represent natural constraints on articulator motion due to physiological limitations and inertia and can reduce ambiguity in the mapping by eliminating unlikely solutions. Continuity constraints can be strictly defined, or loosely enforced using cost functions that penalize discontinuous trajectories.

The inverse problem for the frame-based case can be formulated as follows. Given a sequence of M acoustic feature vectors, $\mathbf{a}_m$, $m \in [1, M]$, extracted from an utterance, calculate I, the sequence of indices of the linked-codebook Φ that best fits the acoustic sequence, $\mathbf{a}$, according to some cost function, $D_{total}(\mathbf{a}, I)$.

As proposed by Sondhi and Schroeter [25], an appropriate cost function is one that minimizes the overall acoustic error and limits the possible solutions to those that vary smoothly. The cost function, $D_{total}(\mathbf{a}, I)$, is typically expressed as the combination

of two costs: one measuring the acoustic distance between the original utterance and its resynthesis and the other measuring continuity in the articulatory parameters or some articulatory representation.

$$D_{total}(\mathbf{a}, I) = D_{acoust}(\mathbf{a}, I) + D_{artic}(I) \tag{5.2}$$

Generally for the frame-based case, D_{acoust} is merely the sum of the acoustic distances in each frame.

$$D_{acoust}(\mathbf{a}, I) = \sum_{m=1}^{M} d_{acoust}(\mathbf{a}_m, I_m) \tag{5.3}$$

$D_{artic}(I)$ is a function of the entire articulatory trajectory. Often, it is convenient to define D_{artic} as a summation of the smoothness contributions, d_{artic}, of each frame. Each frame's contribution represents smoothness in the local region.

$$D_{artic}(I) = \sum_{m=1}^{M} d_{artic}(I, m) \tag{5.4}$$

Finding the optimal trajectory according to the composite metric of Equation (5.2) can be achieved efficiently using a procedure called dynamic programming [60]. Dynamic programming is an efficient technique for finding an the optimal path through a set of nodes, according to some path metric. The path metric is a combination of two measures: one that measures the cost of traveling between points(transition costs) and one that measures the cost of passing through points(node costs). In our problem, we seek the best articulatory path that minimizes Equation (5.2). The acoustic distance of each frame is the node cost, while the articulatory smoothness at a frame is the transition cost.

The result of codebook-lookup followed by dynamic programming is generally not a sufficient final solution due to the coarseness of the codebook sampling. Therefore,

frame-based numerical optimization can be used on the dynamic program result to improve both the acoustic fit and the articulatory smoothness.

Our inverse mapping procedures take advantage of time-frequency/source-tract division within the synthesizer to estimate vocal-tract shape independently of the state of the glottal-source. Independent estimation of glottal-source and vocal-tract parameters cannot properly resolve the interrelated contributions of the source and vocal-tract. But due to the limited accuracy of our synthesizer and articulatory models, and the limited precision of our inverse algorithms, accurate resolution of source and tract contributions is not a practical goal. Therefore, independent estimation of source and tract parameters is used for simplicity in implementation and a reduction in the number of parameters to be estimated simultaneously.

The nature of the goodness measure has a significant effect on the results of the dynamic programming approach. While an optimal path can always be found, this path may not always correspond to a physiologically plausible articulatory trajectory. Additionally, the resynthesis may be unintelligible. The selection of an acoustic feature/distance measure, an articulatory distance measure, and the relative weighting between the two are important research topics. Appropriate solutions depend on the type of speech being processed, the type of mapping algorithm, and the quality of the forward model (synthesizer).

5.2 Inverse Mapping Using Formant Frequencies As Acoustic Features

5.2.1 Procedure

The procedure of linked-codebook lookup and dynamic programming is applied to an inverse mapping problem using the LAM as the articulatory representation and

the first three formant frequencies as the acoustic feature. Both the acoustic and articulatory distances are Euclidean (squared) so that

$$d_{acoust}(\mathbf{a}_m, I_m) = \|(\mathbf{a}_m - \Phi_{ff}(I_m)\|_2^2 \tag{5.5}$$

and

$$d_{artic}(I, m) = \begin{cases} 0, & m = 1 \\ \|\Phi_*(I_m) - \Phi_*(I_{m-1})\|_2^2, & m > 1 \end{cases} \tag{5.6}$$

The test utterance is the question "Why were you away a year Roy?" spoken by a male talker in a quiet room with an inexpensive microphone. The first three formant frequencies, F0, and rms power were extracted for 15 ms frames using the ESPS `formant` routine. Some hand tuning of the resulting formant trajectories was required.

Linked-codebook lookup on a 40000 entry codebook is used to select the $N = 128$ codebook configurations for each frame that minimize the Euclidean distance between the first three formant frequencies of the speech segment and the synthetic transfer function. Dynamic programming uses the same acoustic distance (node cost) along with an articulatory distance (transition cost) defined as Euclidean distance between LAM parameters of adjacent frames. Weighting between the total acoustic distance, d_{acoust}, and the total articulatory distance, d_{artic}, is included in the overall distance measure, D_{total}, by defining a weighting factor, γ $0 < \gamma < 1$, so that

$$D_{total} = \frac{\gamma}{\bar{D}_{acoust}} D_{acoust} + \frac{1-\gamma}{\bar{D}_{artic}} D_{artic}, \tag{5.7}$$

where $\bar{D}_{acoust}$ and $\bar{D}_{artic}$ are normalizing factors depending on the acoustic and articulatory feature/distance measures respectively. As γ is increased, the relative weighting of acoustic distance increases over that of articulatory distance. Without normalization by the average distances, $\bar{D}_{acoust}$ and $\bar{D}_{artic}$, a γ of 0.5 would not result in equal

contributions of articulatory and acoustic distances in most cases, since $\bar{D}_{acoust}$ and $\bar{D}_{artic}$ can differ by orders of magnitude. Since we will be experimenting with a variety of acoustic and articulatory features, it will be useful to have a weighting factor within dynamic programming that is normalized. Average acoustic distance, $\bar{D}_{acoust}$, can be estimated as the average of the acoustic distances of the N-best candidates for all frames of a test utterance. A crude estimate of the average articulatory distance, $\bar{D}_{artic}$, can be calculated by averaging the articulatory distance between the best 5–10 codewords between adjacent frames for all frame boundaries. More complex, time-varying weightings have been reported [4] and may be necessary for utterances containing more than just voiced sounds.

Iterative numerical optimization can improve upon the results of codebook-lookup and dynamic programming by reducing the acoustic distance for each frame and/or reducing the transition costs between frames. Our optimization strategy is as follows.

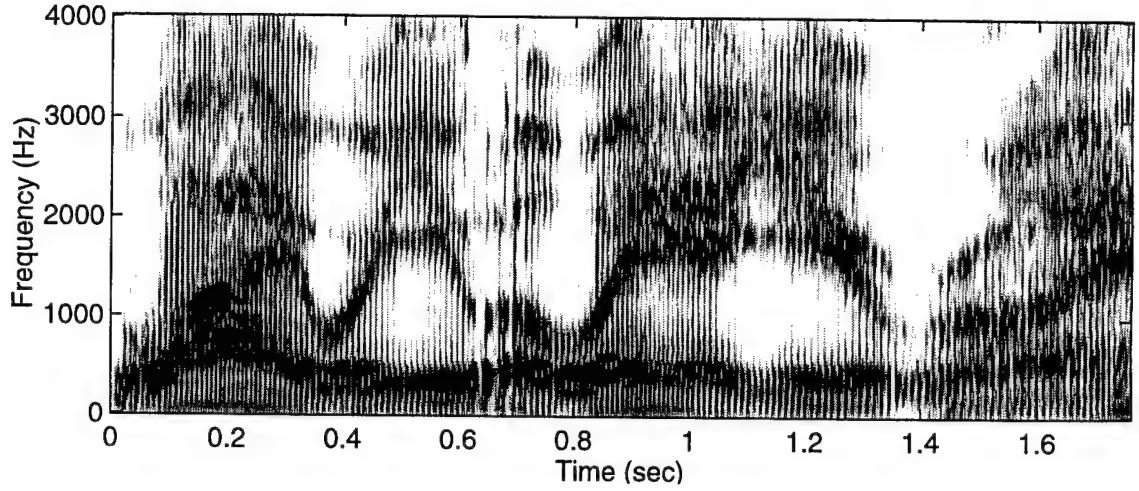
1. Reduce the maximum formant frequency deviation (of the first three formants) to less than 3% for all frames.

2. For each frame, minimize the total transition cost between the preceding and following frames, constrained to keeping the maximum formant frequency deviation to less than 3%.

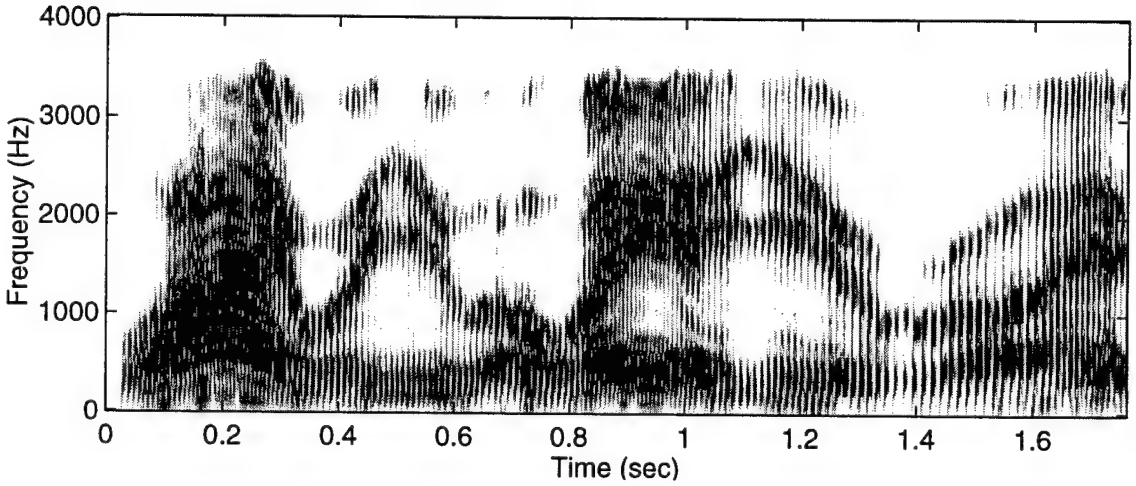
3. Repeat 2 as necessary to increase smoothness.

5.2.2 Results

Figure 5.1 shows the spectrogram of the original utterance and the resynthesized version after codebook-lookup, dynamic programming, and iterative optimization. Formant tracks match closely and there is no observable transient energy due to poor



(a) Original



(b) Resynthesized

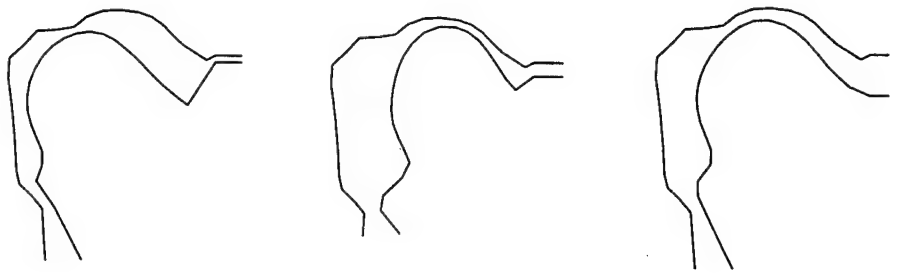
Figure 5.1: Spectrograms of the original and resynthesized versions of the utterance "Why were you away a year Roy?".

	F1	F2	F3
MAX (Hz)	17	53	79
AVG (Hz)	9	25	33

Table 5.1: Maximum and average error in formant frequency estimates for the first three formants.

frame transitions. The desired and estimated formant trajectories are very close with average and maximum frequency deviations as shown in Table 5.1. The close match in formant trajectories makes the resynthesis quite intelligible, but unfortunately, the resynthesis is not completely natural. More effort needs to be spent on the source model and parameters to improve naturalness. Figure 5.2 shows LAM cross-sections at different points along the utterance. Most of the configurations are reasonable, although the LAM has a tendency to form unnatural constrictions in the pharyngeal region. This tendency has also been observed by Boë et al. [39] who suggest that it may be avoided with appropriate codebook pruning. The estimated configurations for [r] in "were" and "year Roy" were bunched rather than retroflex, which is consistent with most speakers of American English.

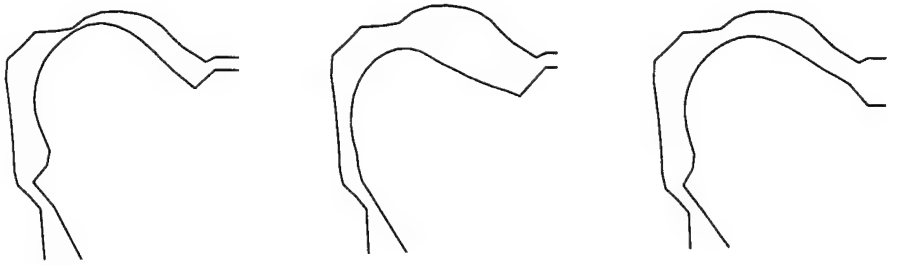
Instead of merely selecting the best acoustic match from the linked-codebook for each frame, dynamic programming selects a smooth articulatory trajectory out of the top N acoustic matches for each frame. This improvement in articulatory trajectory is obtained at the expense of an increased overall acoustic distance as shown in Figure 5.3. The tradeoff within dynamic programming between acoustic fitness and articulatory smoothness is controlled by the weighting factor, γ, and has a significant effect on the resulting trajectory. With formant frequencies as our



(a) Frame 26, [w] of "were"

(b) Frame 36, [j] of "you"

(c) Frame 64, [ɛɪ] of "away"

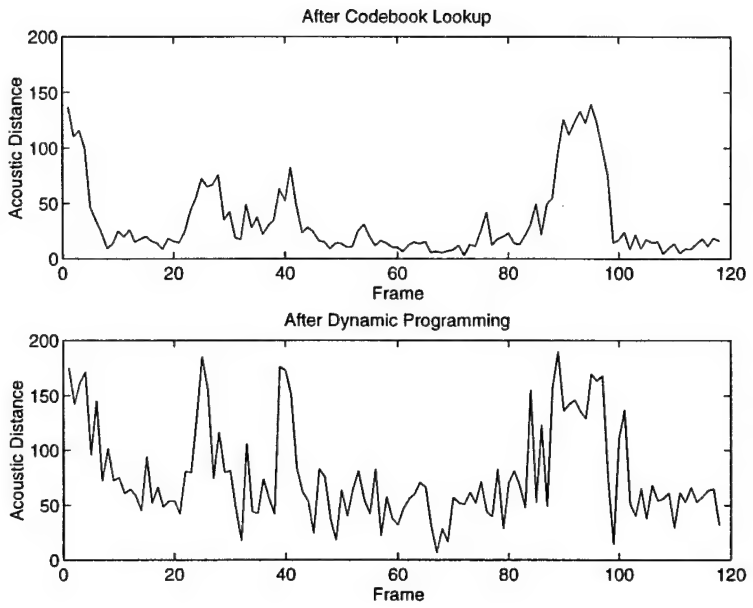


(d) Frame 88, [ɚ] of "year"

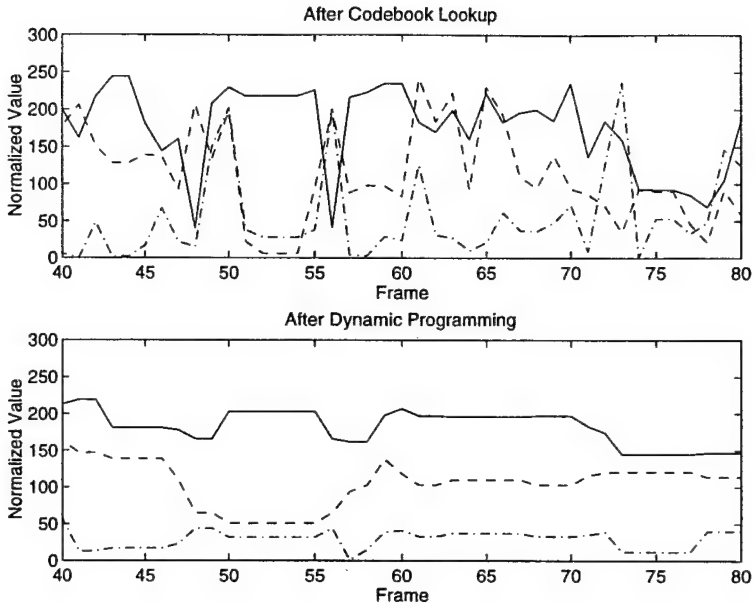
(e) Frame 107, [oɪ] of "Roy" (onset)

(f) Frame 115, [oɪ] of "Roy" (offset)

Figure 5.2: Selected LAM configurations from the acoustic-to-articulatory mapping of the utterance "Why were you away a year Roy?".



(a) Euclidean distance between estimated and actual first three formant frequencies



(b) Articulatory trajectory of three LAM parameters (normalized to [0,255]): tongue-body shape(solid), tongue-tip position(dashed), larynx height(dash-dotted).

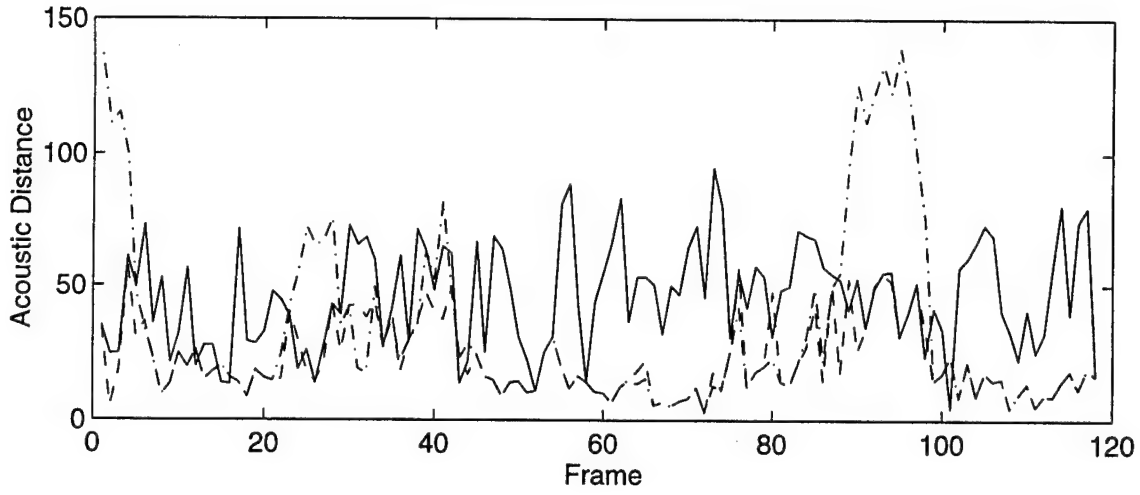
Figure 5.3: Comparison of acoustic fitness and articulatory smoothness before and after dynamic programming.

acoustic feature, most of the top N codewords are reasonable fits so acoustic distance is weighted lightly. We found that 0.1 was a reasonable choice for γ in this case.

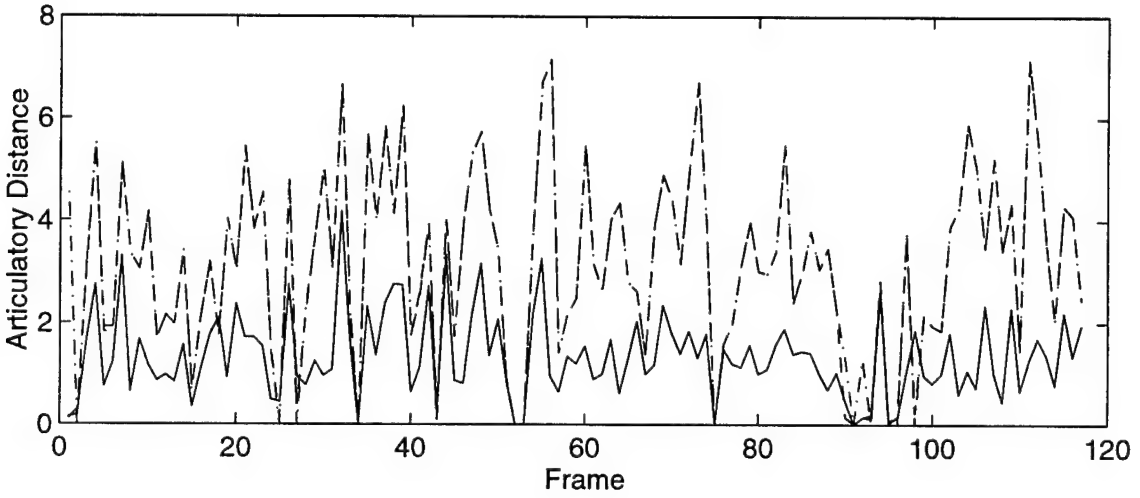
The effects of the two stages of optimization are shown in Figure 5.4. The first optimization improves the acoustic fit of each frame, without regard to trajectory smoothness. Its main purpose is to correct poor acoustic fits due to gaps in the codebook. The biggest improvements occur in three or four regions, typically [r] and [w] sounds, where the codebook could not find good matches. The second, smoothing optimization modifies the trajectory without significantly affecting the acoustic fit. The acoustic distance does increase after the second optimization, but is maintained within the 3% constraints. The smoothing optimization improves average frame transition cost by 35%. While this is a significant improvement, the actual configurations are not drastically altered. Figure 5.5 shows the vocal-tract cross-section corresponding to one of the largest changes in the smoothing optimization. The change is not profound; however, there are a few sharp transitions, such as those at frames 32 and 7, that may not reflect natural movement of the articulators. Further smoothing optimization would continue to improve trajectory smoothness, but with a limiting effect. As we see in Figure 5.5, the overall change in the trajectory will be small.

5.2.3 Discussion

While the above technique may not be the perfect solution, it is a working solution that acts as a starting point upon which to improve. We have intelligible resynthesis, acoustic-to-articulatory mapping results with reasonable articulatory configurations and, in most cases, smooth frame transitions. Work on the glottal-source is necessary



(a) Acoustic distance



(b) Articulatory transition distance

Figure 5.4: Per frame acoustic and articulatory transition costs for three stages of acoustic-to-articulatory mapping of the utterance "Why were you away a year Roy?": after dynamic programming(dash-dotted), after first optimization(dashed), after smoothing optimization(solid). Note that articulatory transitions after dynamic programming and after the first optimization are nearly identical.

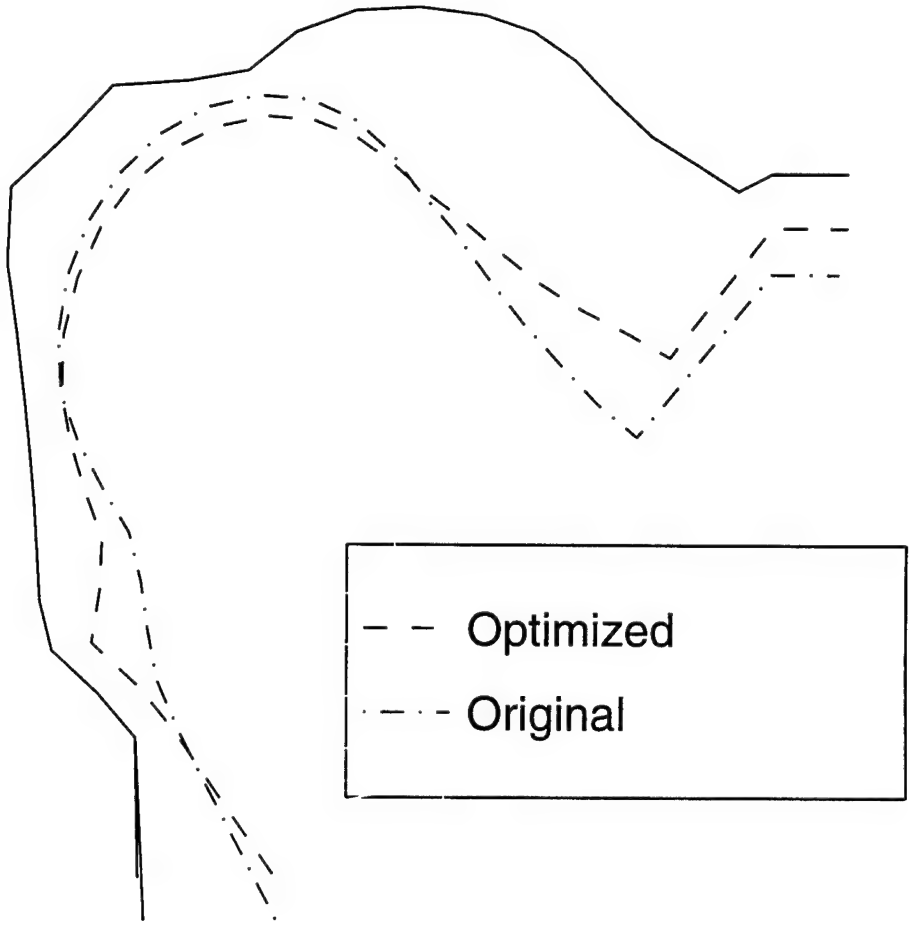


Figure 5.5: Articulatory configurations for frame 124, in which one of the larger articulatory changes was made during the two optimization stages. The dashed line corresponds with the optimized trajectory; the dash-dotted line corresponds to the configuration before optimization.

to improve naturalness, but that is somewhat independent of the vocal-tract estimation. Some comments and observations can be made on the algorithm in general and the use of formants as an acoustic feature.

It does not appear to be difficult to find articulatory trajectories such that the first three formant frequencies are within 3% of their desired value. The more difficult problem is selecting a trajectory that is reasonable, in terms of smoothness and production plausibility. The approach of linked-codebook lookup and dynamic programming quickly provides us with a reasonable "global" solution that can be fine-tuned using iterative optimization. It is important to note that iterative optimization does not drastically alter the overall articulatory trajectory. Dynamic programming solves the majority of the inverse transformation, with iterative optimization merely ensuring an acceptable resynthesis. Therefore, if the dynamic programming result includes an unnatural transition, possibly due to the coarseness of the linked-codebook, frame-based iterative optimization cannot detect or correct the problem. This suggests an alternative approach to the inverse mapping problem that places more responsibility on iterative optimization. By starting iterative optimization from an adjacent frame configuration, proximity in articulatory and acoustic domains can be exploited to find solutions that produce good acoustic matches and provide very good frame-to-frame transitions. This approach has been used in a number of acoustic-to-articulatory mapping algorithms [3, 61, 62]. This approach does not offer the same "global" optimization perspective that dynamic programming does, but offers smoother "local" trajectories. The comparison and, perhaps, combination of these two approaches is a fertile area for future investigation.

Are formants good features for acoustic-to-articulatory mapping of voiced sounds? Formant frequencies, or their logarithms, have been used in many acoustic-to-articulatory mapping studies [24, 12, 63, 16, 20, 9]. Our positive results suggest that formant frequencies are strong features, but it is likely their performance can be improved. Formants are a perceptually salient feature of voiced sounds. Their use as an acoustic feature ensures intelligible resynthesis, which is a necessary condition of our result. For better or worse, a seven dimensional articulatory representation is estimated on each frame from a single three dimensional acoustic feature. The four unconstrained dimensions provide the freedom to match formant frequencies and make adjustments for smoothness. Whether this permits or prevents accurate acoustic-to-articulatory mapping results is not clear, but since formants are such strong features for voiced sounds, we might be "getting away" with this underdetermined situation. Since formant frequencies are just a partial description of entire spectrum, we are not, directly or indirectly, matching formant bandwidths, formant amplitudes, spectral tilt, etc. Clearly, there is much more information in speech signal that is not being used. More acoustic dimensions can eliminate the underdetermined situation and should add information to help improve acoustic-to-articulatory mapping results. But if significant model mismatch exists between actual speech production and our articulatory speech synthesizer due to the assumptions, simplifications, and constraints of our forward model, the extra information might obscure a reasonable solution with acoustic features of secondary importance. For example, Sorokin [24] performed static inversion using log formant frequencies and compared results to X-ray microbeam data. He found that results can worsen when four formants are used over three and observed, "the same situation emerges when amplitudes of formants are included in the acoustic

constraints – in many cases, formant and tongue shape matching get worse instead of better." The best compromise between underdetermined and over-constrained features is likely to depend on the forward model, the inversion technique, and the ultimate goal of the acoustic-to-articulatory routine.

While formant frequencies have provided us with a working solution, there are a number of reasons why formant frequencies are not good for our applications. First, the acoustic-to-articulatory mapping routine is very sensitive to incorrect formant frequencies, therefore hand inspection and tuning of the automated formant extraction results is required. Second, formant frequencies are not good features for non-voiced sounds such as fricatives. While we may use different features for voiced and fricated sounds, it may be desirable to use the same, or similar acoustic features for both fricated and non-fricated portions. Therefore, we need to investigate using acoustic features other than formant frequencies for our acoustic-to-articulatory mapping algorithm.

5.3 Inverse Mapping Using Alternative Acoustic Features

There is a large variety of speech features which could be applied to the acoustic-to-articulatory mapping problem. Most features in the literature are either PSD-based, LPC-based, or FFT cepstrum-based. Perceptual weighting or frequency scaling can be applied to any of the features as well. We have chosen to investigate cepstral-based features due to their potential as features for fricatives as well as their demonstrated utility for voiced sounds [64, 30, 61]. Another strong impetus for their use is that cepstral features, with appropriate weighting, have been shown to offer reduced sensitivity to variations in the glottal-source. Weighted FFT cepstral coefficients with

weighting as defined by Meyer, Schroeter and Sondhi [30] was used as an alternative feature in our acoustic-to-articulatory mapping algorithm. This weighting is designed specifically to reduce source influence in a acoustic-to-articulatory mapping routine.

$$wgt(l) = \begin{cases} (\frac{l}{20})^{0.4}, & l = 1 \ldots 20 \\ 0.5 + 0.5 \cos \pi \frac{l-20}{20}, & l = 21 \ldots 30 \end{cases} \tag{5.8}$$

5.3.1 Results and Discussion

Linked-codebook lookup, dynamic programming, and iterative optimization, was applied to the utterance "Why were you away a year Roy" using the cepstral feature. It did not produce produce acceptable results. Figure 5.6 shows formant trajectories after dynamic programming with $\gamma = 0.7$ for a typical case using weighted FFT cepstral coefficients. Dynamic programming is unable to disambiguate close formants, resulting in discontinuous formant trajectories due to skipped formants. Adjusting the weighting between acoustic and articulatory costs within the dynamic programming cannot overcome the problem. Errors occur in the same place for both cepstral features: at the [w] of "why" and the [r] of "year Roy". The location of these errors do not appear related to signal power, except, perhaps, at the start of the utterance where signal power was low. It is interesting to note that some of those frames in which dynamic programming made incorrect decisions are also frames in which the ESPS `formant` routine made errors and had to be hand corrected.

Frame-based iterative optimization is unable to overcome the incorrect choices of dynamic programming. While transition costs are a part of the optimization criterion, frame-based operations cannot drastically change the overall trajectory. Frame-based optimization easily falls into a local minima and must start close to a good solution

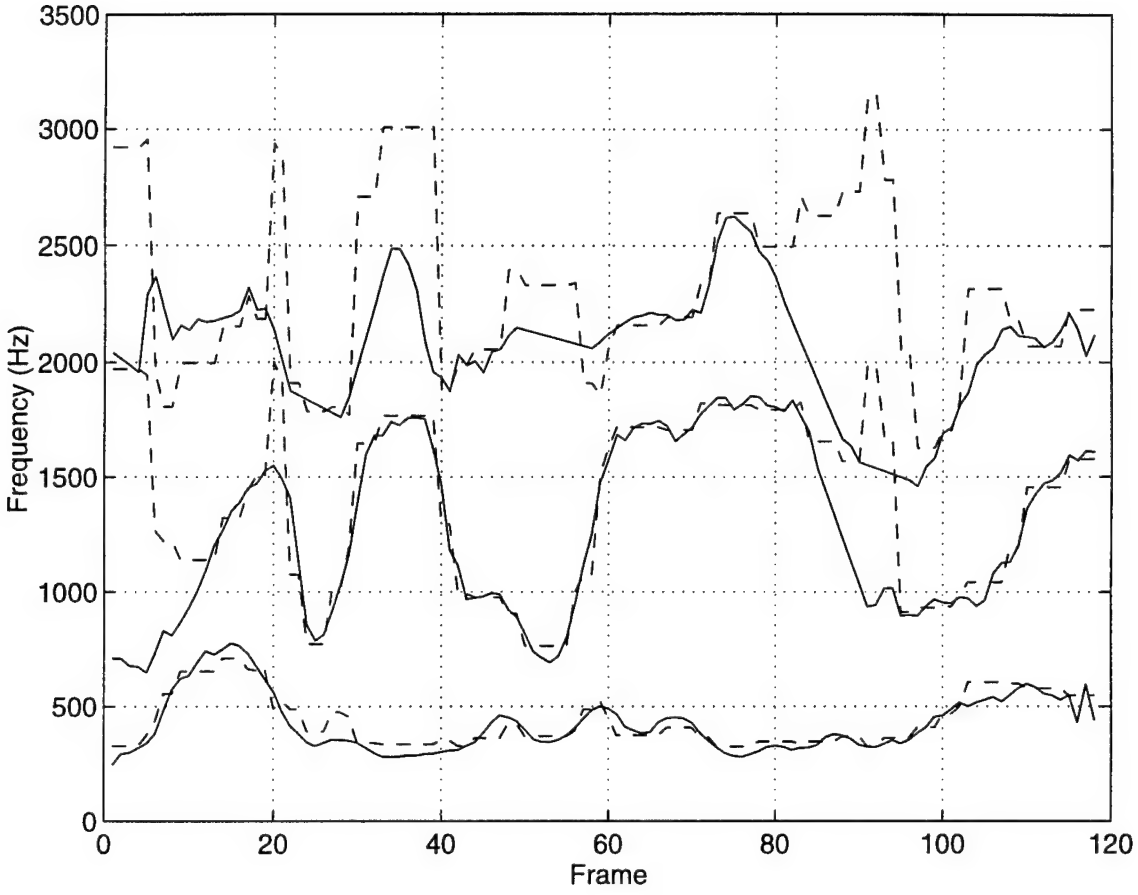


Figure 5.6: Desired (solid) and estimated (dashed) formants trajectories for the utterance "Why were you away a year Roy?" after linked-codebook lookup and dynamic programming. The cost function minimized was a combination of a weighted FFT cepstral acoustic distance and a LAM articulatory distance. Note that further frame-based iterative optimization could not recover from the incorrect formant estimates.

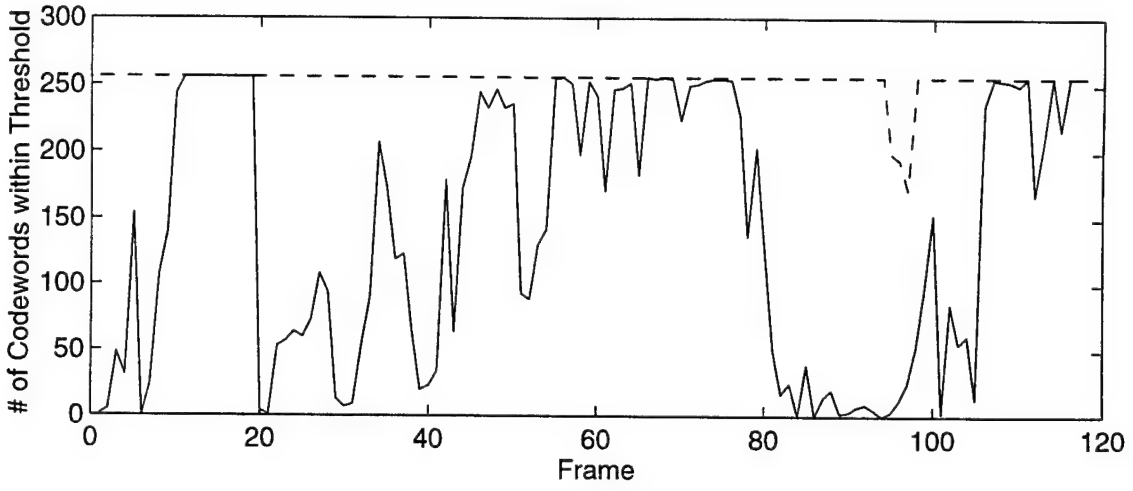


Figure 5.7: The number of codewords in each frame after 256-best linked-codebook lookup whose formant frequencies are within 400 Hz of the true formant frequencies. Linked-codebook lookup used Euclidean distance between weighted FFT cepstral coefficients (solid) and formant frequencies (dashed).

to perform well. Therefore, dynamic programming remains the key for obtaining an acceptable trajectory.

Why can dynamic programming not prune out such poor decisions such as mistaking two close formants as one or skipping over a formant? A close examination of the configurations provided by linked-codebook lookup reveals part of the answer. Figure 5.7 shows a measure of the number of "good" configurations that linked-codebook lookup is producing in each frame. A "good" configuration is one that has formant frequencies near their true values. More "good" configurations gives dynamic programming more choices to work with and reduces the chance that a poor trajectory will be chosen. In a majority of the frames, most of the codewords are good configurations, but clearly there are frames for which linked-codebook lookup produces very few good configurations. Figures 5.6 and 5.7 suggest that the number of good

linked-codebook lookup configurations are small for any frame in which two formants are close together. Therefore, it is no surprise that poor results in dynamic programming occur at or near these frames. This is a very consistent trend that could be attributed to sampling problems/synthesizer characteristics or weaknesses in the acoustic feature/distance metric. Figure 5.7 also shows the number of good configuration available when a formant codebook is used. Except for one location near frame 95 , the [r] of "Roy", there is plenty of good configurations. It is interesting to note that in Figure 5.3(a), which shows the acoustic fitness of acoustic-to-articulatory mapping using formant frequencies, there are a few segments in which the acoustic fitness is much worse than for other frames. In most cases, these segments coincide with the underrepresented segments in Figure 5.7.

5.3.2 Continuity in Resonance

Over certain segments, dynamic programming with cepstral features is unable to disambiguate close formants. Ideally, the continuity component of the cost function should penalize the bad breaks that accompany two formant tracks becoming one, and one becoming two. In practice, dynamic programming is falling into bad local minima due to the paucity of candidates with a correct formant structure. Our experience suggests that frame-based iterative optimization does not drastically alter the overall articulatory trajectory except to smooth it and improve the acoustic fit. Therefore, it is essential that the results of dynamic programming have continuous formant traces. One way to encourage this is to incorporate a formant continuity measure, d_{reson}, into the transition costs within dynamic programming.

$$d_{reson}(i,m) = \begin{cases} 0, & m = 1 \\ \|\Phi_{ff}(I_m) - \Phi_{ff}(I_{m-1})\|_2^2, & m > 1 \end{cases} \tag{5.9}$$

The articulatory distance, d_{artic}, is redefined to be the weighted sum of the original articulatory distance, d_{artic}^{old}, and the formant transition measure, d_{reson}.

$$d_{artic} = \beta d_{artic}^{old} + (1 - \beta) d_{reson} \tag{5.10}$$

The weighting term, β, must be tuned appropriately and can be normalized much like the γ term in Equation (5.7). Unlike formant frequency estimation from real speech, formant frequencies in synthetic spectra are easy to estimate. Resonances are clearly defined in the synthetic transfer function, and are not obscured by glottal-source effects and noise. Finally, to increase the number of candidate configurations with a desirable formant structure, the number of candidate configurations produced by linked-codebook lookup was increased from 128 to 1024.

Figure 5.8 shows the formant trajectories estimated with dynamic programming using formant continuity in the cost function. Weighting factors, γ and β, were 0.85 and 0.0 respectively. Formant continuity corrects many of the errors experienced in Figure 5.6.

Further iterative optimization of the dynamic programming result, including formant continuity in the cost function, produces reasonable resynthesis and articulatory trajectories with quality approaching that of the results using formants as features (See Figure 5.9. The only area of difficulty is the [r] of "year Roy" where dynamic programming provided a poor starting point. Resonance continuity significantly improves the quality of the resynthesis, purely because it enforces smooth formant trajectories. A big issue in the iterative optimization phase is the appropriate weighting of acoustic verses articulatory distances, and the weighting of articulatory model constraints and formant smoothing constraints. Correct values for these parameters is not as clear as in the dynamic programming case.

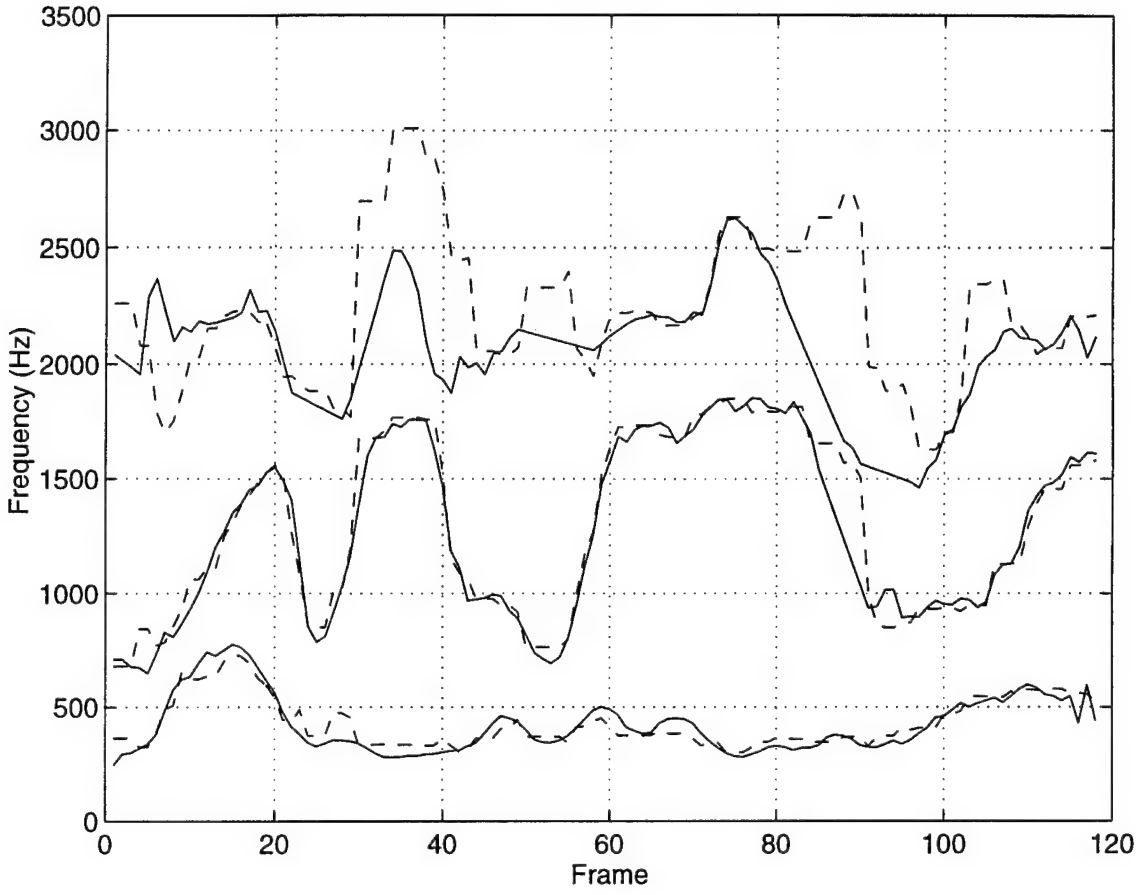


Figure 5.8: Desired (solid) and estimated (dashed) formant trajectories for the utterance "Why were you away a year Roy?" after linked-codebook lookup and dynamic programming. The cost function minimized was a combination of a weighted FFT cepstral acoustic distance and formant smoothing. This is a significant improvement over Figure 5.6.

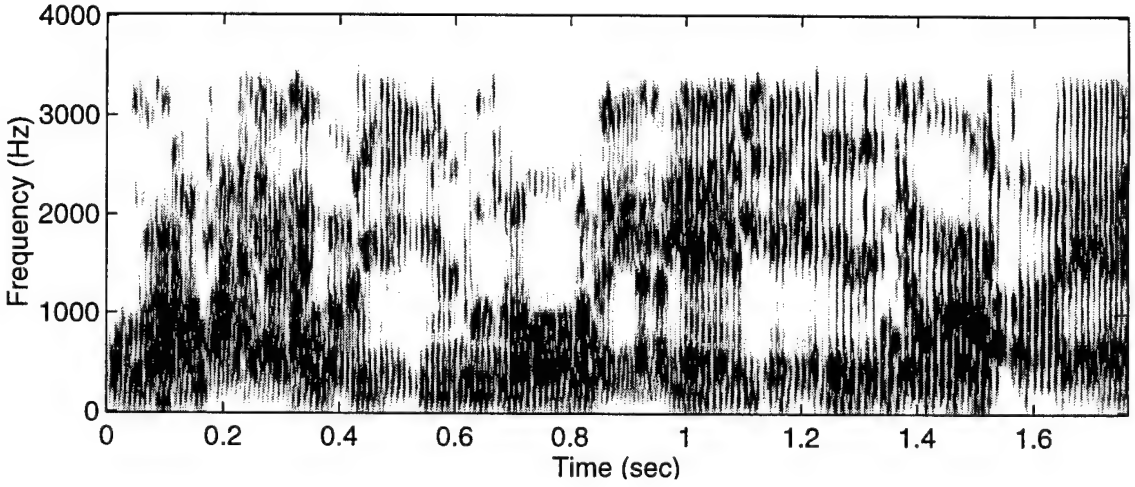


Figure 5.9: Spectrograms of the resynthesized versions of the utterance "Why were you away a year Roy?" after dynamic programming and iterative optimization using weighted FFT cepstral coefficients as the acoustic feature, and resonance continuity in the articulatory distance.

CHAPTER 6

INTEGRATION OF VOICED AND FRICATED SPEECH IN AN INVERSE MAPPING SCHEME

6.1 Introduction

In Chapter 4, fricative linked-codebooks were generated and used to identify effective acoustic features for fricative acoustic-to-articulatory mapping. Ways in which fricative linked-codebook lookup must be constrained to produce reasonable inverse mapping results were identified. These results are used in this chapter to integrate fricatives into the linked-codebook/dynamic programming algorithm of Chapter 5.

The differences between fricatives and vowels have a significant effect on the way acoustic-to-articulatory mapping is performed for each. The same acoustic feature and acoustic distance metric cannot be used for all sounds, since the perceptually significant differences between phones of one class may be different than those for another phonetic class. For example, in Chapter 4, fricative features were designed to be insensitive to the presence of voicing and represent frequencies up to 8 kHz, while the vowel features were chosen for their insensitivity to glottal-source variation and represent frequencies only up to 4 kHz. Similarly, good articulatory distance measures for one class of sounds may not be effective for another class of sounds. Fortunately, our fricative articulatory model is identical to the vowel articulatory model, except

for the addition of an extra parameter specifying frication source location. Therefore, differences in articulatory model-based distance measures may be less pronounced. Changes in phonetic class during acoustic-to-articulatory mapping requires the effective transition between different acoustic and articulatory distance measures. For example, the use of a weighted articulatory distancep-based on the relative "significance" of articulatory dimensions as suggested in Chapter 4 imposes an articulatory distance that is dependent on phonetic class.

In addition to the above articulatory and acoustic differences, there are differences between the acoustic-to-articulatory mapping of fricated and voiced speech that are especially important in the dynamic case. Significant source-tract interaction exists in fricative production. The presence and amplitude of frication is strongly dependent on the area of the constriction. At the same time, since constriction area is small, transfer functions are also quite sensitive to constriction area. The non-linear relation between constriction area and the glottal-source source settings determines frication amplitude as well as the presence and magnitude of voicing. In voiced fricatives, energy is present from two sources: the frication noise pressure source, and the glottal-source. The contribution of each to the speech waveform must be separated in some way in order to estimate vocal-tract shape and balance the relative contribution of each source. The presence of such mixing suggests the need to simultaneously estimate two acoustic transfer functions.

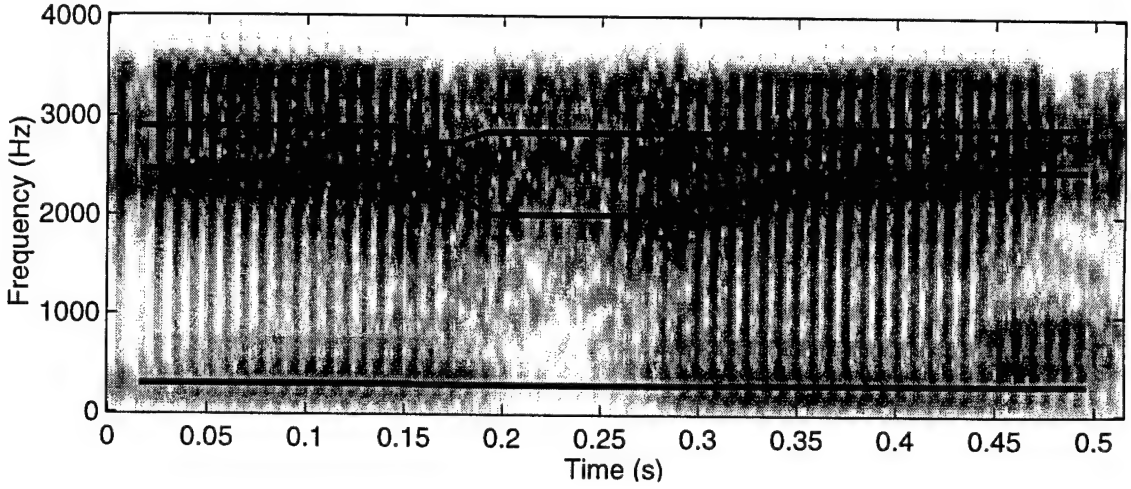
In order to lead up to an acoustic-to-articulatory mapping algorithm for voiced and fricated speech, we present techniques for using contextual information to improve inverse mapping estimates for both vowels and fricatives. These results are

then combined with the results of Chapters 3 through 5 to produce an acoustic-to--articulatory mapping algorithm for voiced speech containing intervocalic fricatives. We discuss source-tract interaction in fricatives and present a heuristic to decouple the estimation of source and vocal-tract parameters. Operation of the algorithm is demonstrated for a variety of vowel-fricative-vowel tokens.

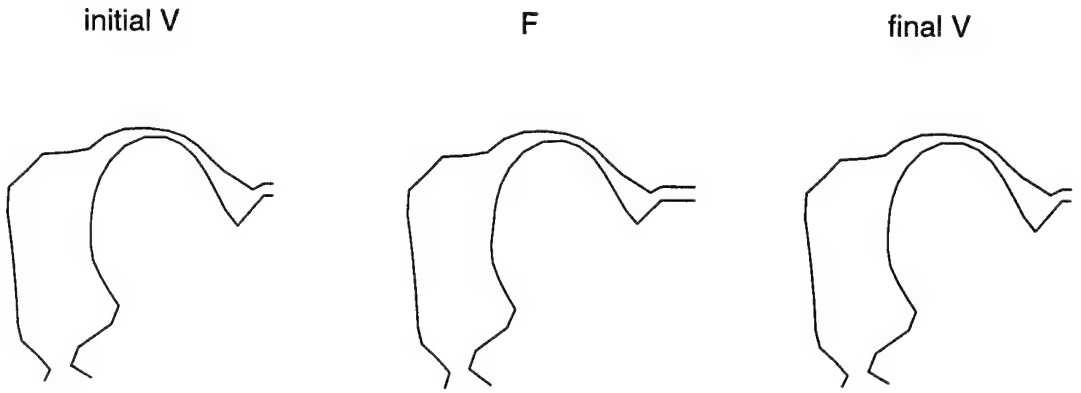
6.2 Contextual Information for Voiced and Fricated Speech

In the frame-based formulation of the acoustic-to-articulatory mapping problem, an articulatory configuration is estimated from a single frame of speech. This estimate can be improved with the use of contextual information, in the form of acoustics and articulatory estimates from adjacent frames. This improvement from contextual information is achieved in the voiced algorithm of Chapter 5 by using dynamic programming to enforce continuity constraints across frames. The additional information helps to alleviate the effect of the many-to-one mapping problem by offering additional information with which to select from possible solutions. Acoustic-to-articulatory mapping of fricated speech should similarly benefit from contextual information, perhaps even to a greater degree, since there appears to be a greater amount of uncertainty in static fricative estimates.

Formant transitions into and out of fricatives often show rapid changes. These dynamics are cues for fricative perception and may be used to indicate the direction in which the articulators are moving. In many cases, these formants are visible well after the start of frication so a voiced acoustic-to-articulatory mapping should be able to extract some information from this segment. For example, Figures 6.1 through 6.4 show the results of directly applying the voiced, dynamic acoustic-to-articulatory

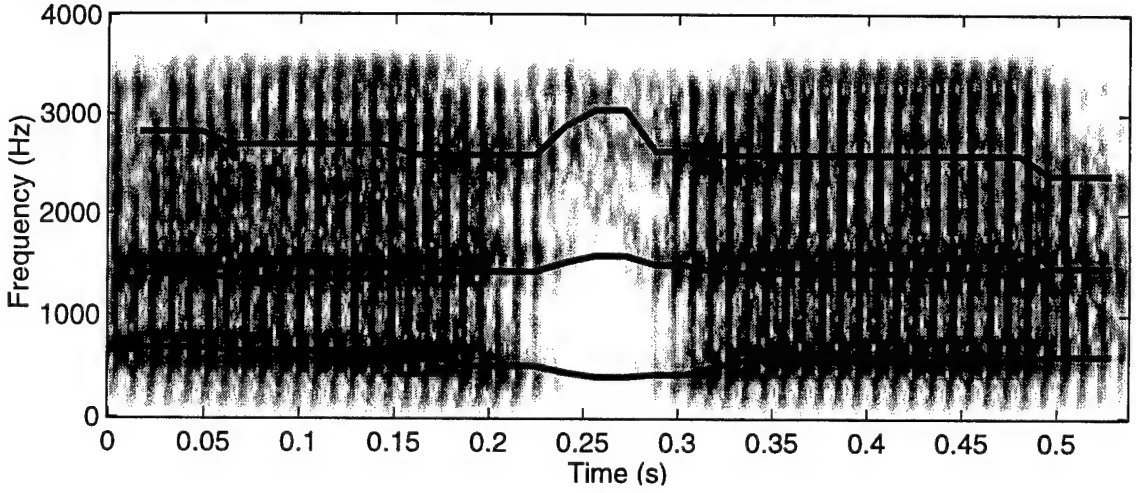


(a) Spectrogram and estimated formants.

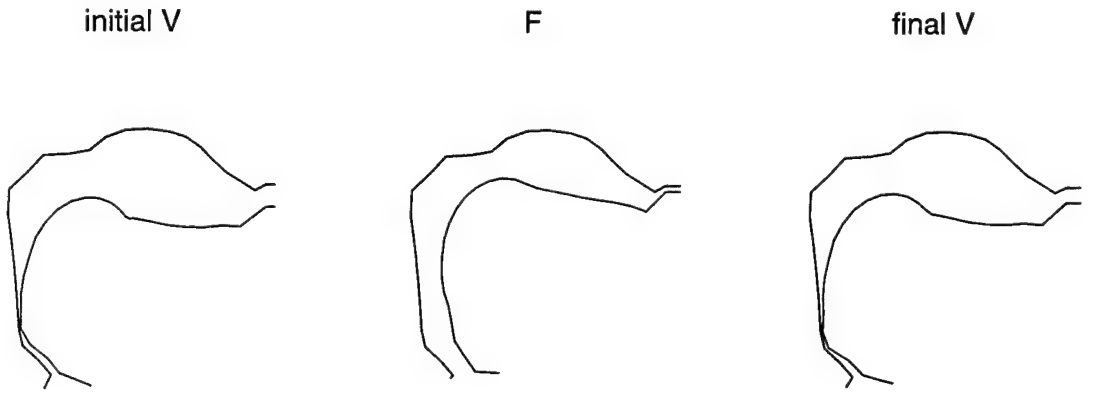


(b) Central articulatory configuration of each segment.

Figure 6.1: Results of applying voiced, dynamic acoustic-to-articulatory mapping algorithm to /izi/.

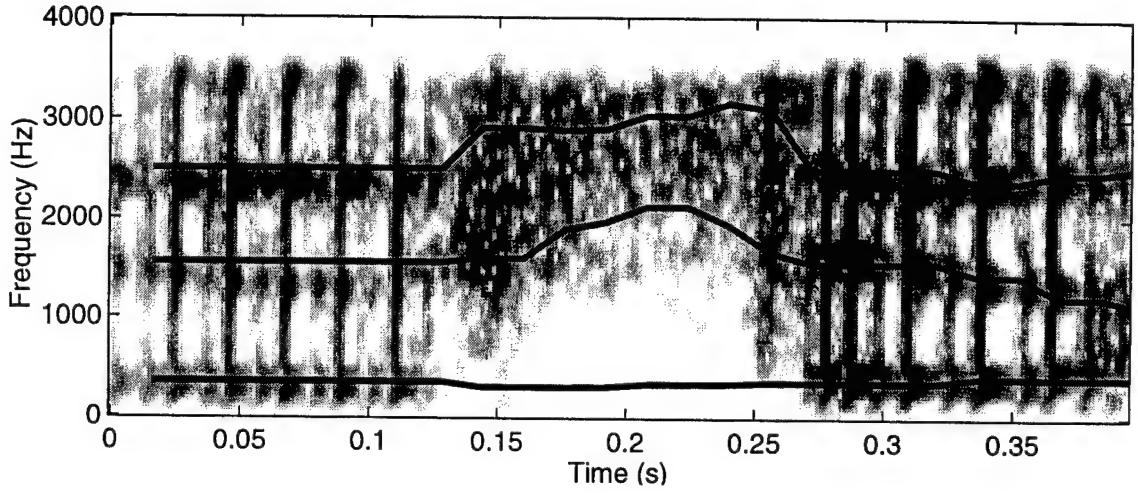


(a) Spectrogram and estimated formants.

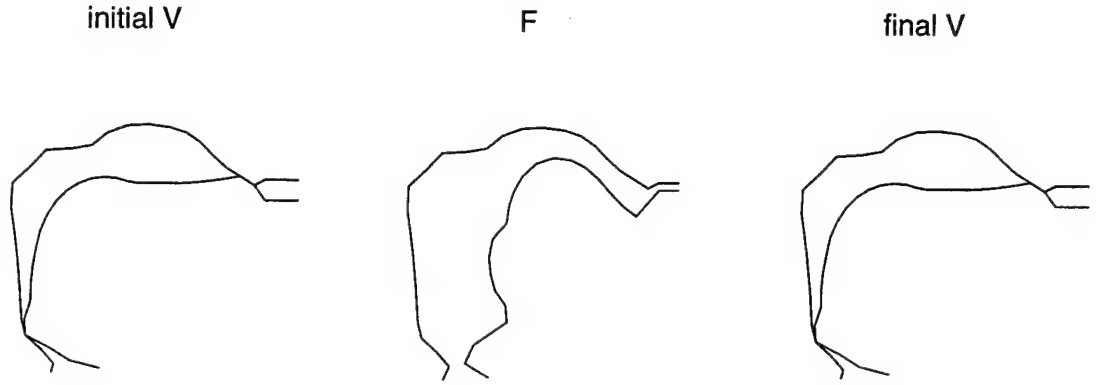


(b) Central articulatory configuration of each segment.

Figure 6.2: Results of applying voiced, dynamic acoustic-to-articulatory mapping algorithm to /ɑzɑ/.

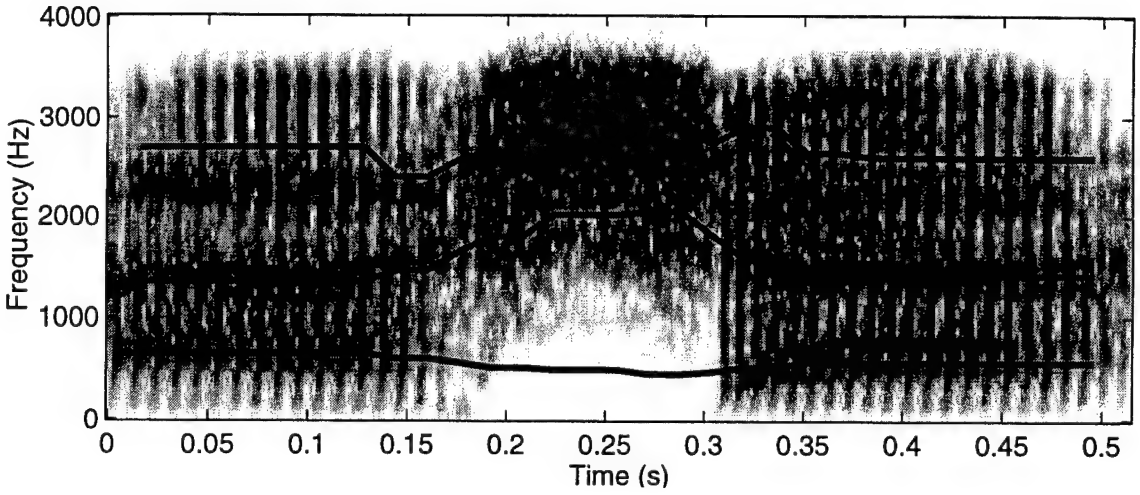


(a) Spectrogram and estimated formants.

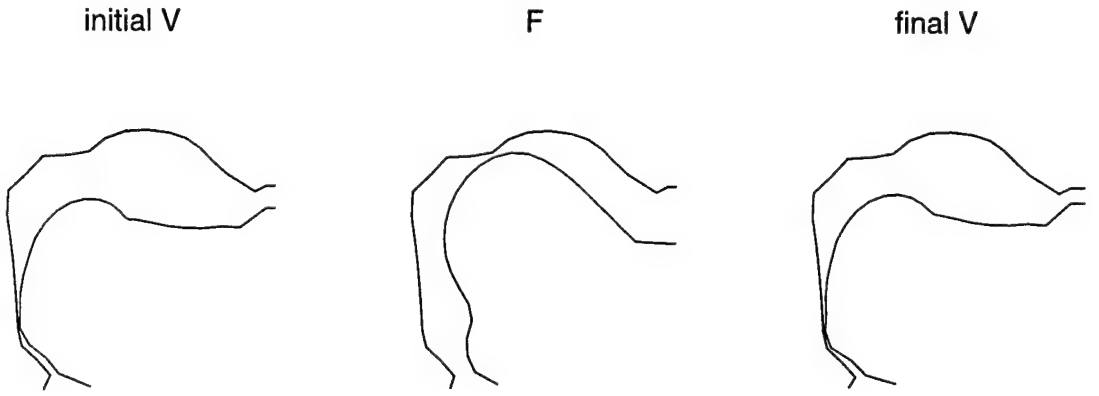


(b) Central articulatory configuration of each segment.

Figure 6.3: Results of applying voiced, dynamic acoustic-to-articulatory mapping algorithm to /aθa/.



(a) Spectrogram and estimated formants.



(b) Central articulatory configuration of each segment.

Figure 6.4: Results of applying voiced, dynamic acoustic-to-articulatory mapping algorithm to /ɑʃɑ/.

mapping routine to speech containing intervocalic fricatives. In the fricative segments, the acoustic-to-articulatory mapping results do not merely hold the same configuration in the absence of a visible formant structure in the spectrogram. Often, the formant trajectories are continued after the formant structure disappears as can be seen for the voiced fricatives in Figures 6.1 and 6.2. This suggests that the algorithm may be tracking some lower amplitude formant energies at the fricative boundaries resulting from some limited vibrations that still remain in the glottis. For the unvoiced fricatives in Figures 6.3 and 6.4, some change in the articulatory estimates can be observed during the fricative segments, but it seems that the algorithm may be tracking frication energy. This is quite apparent for palato-alveolar fricatives, which have high amplitude frication as low as 2.5 kHz which overpower any low level formant energy. One way in which the tracking of fricative energy may be reduced is by performing acoustic-to-articulatory mapping that ends part way into the fricative and starts back up again near the end of the fricative. This approach is suggested in [4] to deal with the acoustic-to-articulatory mapping of silences gaps in stops. How this would improve results remains to be investigated.

In most cases, the acoustic-to-articulatory mapping produces vowels with reasonable configurations and good formant locations. The presence of the central fricative does not appear to adversely effect the vowel inverse mapping. Of the 48 VFV tokens spoken by speaker MJC, only those for /u/ with palato-alveolar fricatives made incorrect choices for F1 and F2. Apparently, the LAM has difficulty producing this vowel consistently.

6.2.1 Experiment

Since the voiced acoustic-to-articulatory mapping algorithm may be tracking some low-level formant energy during portions of the fricative segment, we would like to see if that information can be used to improve our fricative estimates.

Voiced dynamic acoustic-to-articulatory mapping was performed on VFV tokens to generate contextual information in the form of an estimated articulatory feature, **p**, for the central fricative frame. The articulatory feature is considered contextual due to the dynamic programming enforcement of continuity in both the vowel and fricative segments. For static fricative acoustic-to-articulatory mapping this contextual information can be incorporated into the lookup by adding to the acoustic distance measure, d_{acoust}, a measure of articulatory distance, d_{geo}, from the contextual feature, **p**.

$$d_{total} = d_{acoust} + \delta d_{geo} \tag{6.1}$$

The contribution of the articulatory distance is weighted by δ so that as δ is increased from zero, the contribution of contextual information increases.

Since the true fricative articulatory shape is not known for the VFV tokens, it is difficult to find an objective measure of the effect of contextual information on fricative inverse mapping. From the results of Chapter 4, we know that constrictions of non-sibilant fricatives (/f/, /v/, /θ/, /ð/) are best located at the lips while constrictions for sibilant fricatives (/s/, /z/, /ʃ/, /ʒ/) are best located posterior to the lips. We can use the placement of the constriction location after static fricative acoustic-to--articulatory mapping to classify the VFV tokens as sibilant or non-sibilant. The classification accuracy is an objective measure of performance.

	$\delta = 0.0$	$\delta = 0.001$	$\delta = 0.01$
Overall	85.42	93.75	68.75
Sibilant	95.83	91.67	37.50
Non-sibilant	75.00	95.83	100.00
Unvoiced	83.33	91.67	79.17
Voiced	87.50	95.83	58.33
/i/	93.75	100.00	68.75
/ɑ/	87.50	93.75	75.00
/u/	75.00	87.50	62.50

Table 6.1: Sibilant/non-sibilant classification accuracy (%) for static acoustic-to-articulatory mapping using distance measures incorporating articulatory distances weighted by δ.

Table 6.1 shows the results of static acoustic-to-articulatory mapping with contextual information for 48 VFV tokens of speaker MJC. Normalized FPSD2 correlation was used as an acoustic distance and LAM Euclidean distance, weighted by δ, was used as an articulatory distance. Without any contextual information, static lookup correctly locates the constriction 85% of the time. The majority of errors occur in mistaking non-sibilants for sibilants. The addition of an appropriately weighted articulatory distance, 0.001 in this case, static lookup performance improves to almost 94%. If the articulatory distance weight is too large, classification performance is degraded. Since we only have 48 tokens, the strength of our conclusions is limited. Nevertheless, the results suggest that contextual information, even used in this simplistic manner, does help fricative acoustic-to-articulatory mapping.

The next section describes a procedure for the acoustic-to-articulatory mapping of continuous speech containing fricatives. The procedure attempts to incorporate contextual information between voiced and fricated frames as demonstrated in this section.

6.3 An Algorithm for Inverse Mapping of Voiced and Fricated Speech

Our procedure for acoustic-to-articulatory mapping of continuous speech containing fricatives contains five steps that perform separate inversion of voiced and fricated segments, followed by re-estimation in the context of the previous estimates, followed by finer adjustments of constriction area and source parameters to produce good resynthesis. The procedure provides an initial estimate of the articulatory trajectory which can be used to seed further iterative optimization. The procedure is illustrated in Figure 6.5 and described in more detail below.

The speech is assumed to be presegmented into voiced and fricated regions. Additionally, for selection of source parameter trajectories and resynthesis, fricatives must be classified as voiced or unvoiced. Again, voiced speech and voiced acoustic-to-articulatory mapping refers to only vowels and glides, without liquids or any fricated, aspirated, nasalized, or plosive sounds.

6.3.1 Step One: Voiced Speech Processing

As illustrated in Section 6.2, direct application of the acoustic-to-articulatory mapping algorithm for voiced, continuous speech to speech containing fricatives produces acceptable results in the voiced segments *as well as* natural looking transitions during portions of the fricative segments. These transitions follow the formant structure visible in many of the boundary regions between vowel and fricative. Therefore, the voiced acoustic-to-articulatory mapping routine of Chapter 5 is the first step in our algorithm. The algorithm uses weighted FFT cepstral acoustic features and an articulatory distance measure combining distance between LAM parameters and

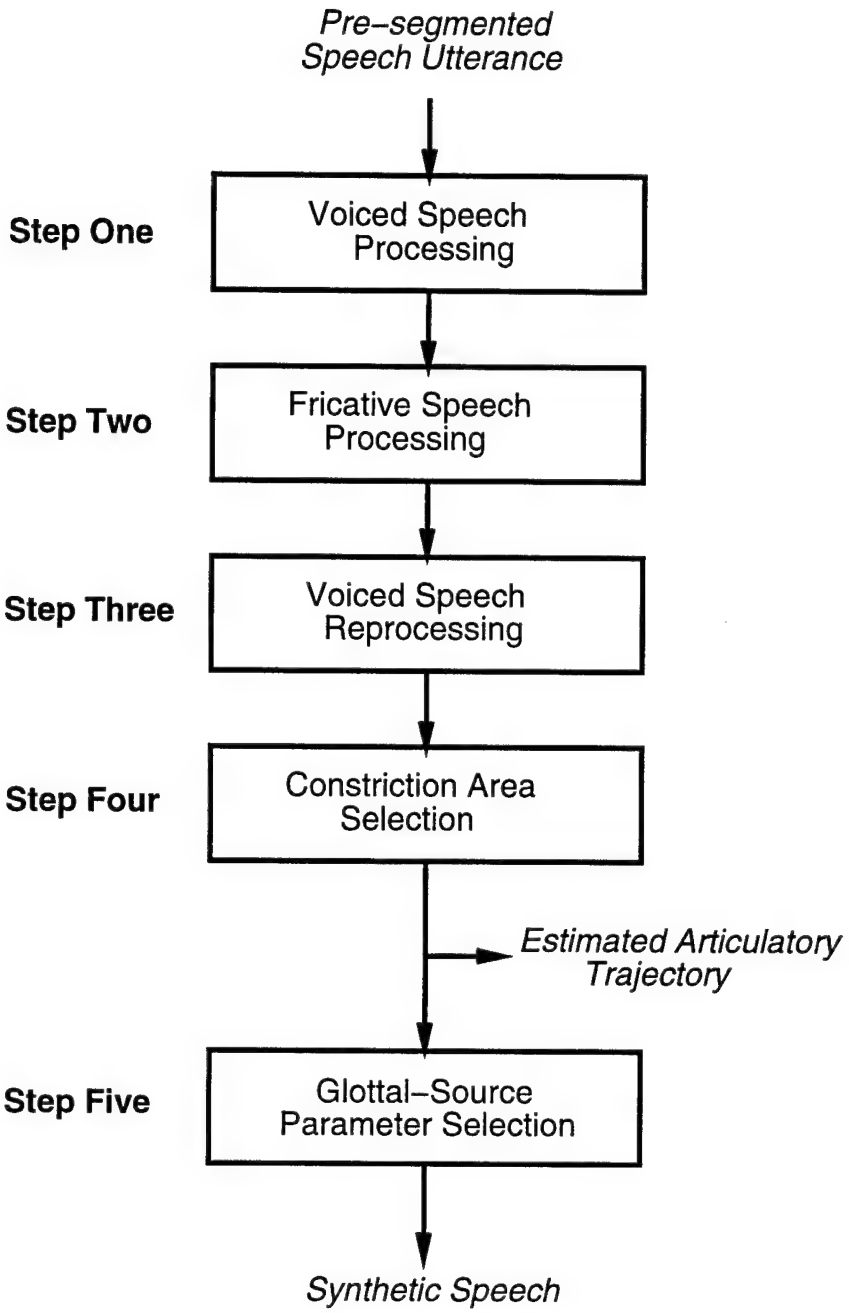


Figure 6.5: Processing flow in acoustic-to-articulatory mapping algorithm for continuous speech containing voiced and fricated speech.

distance between resonant frequencies. These features are used with the same relative weightings as described in Section 5.3.2.

6.3.2 Step Two: Fricative Speech Processing

This step estimates a single articulatory configuration for the central portion of the fricative segment. Fricative linked-codebook lookup is performed as described in Chapter 4 using the FPSD2 acoustic feature and the FLAM model. The linked-codebook is constrained in constriction location and frication location as suggested in Chapter 4. Additionally, constriction area is limited to a small range (0.1–0.15 cm^2) for which our predefined source parameter trajectories are designed to operate. The distance measure used in linked-codebook lookup includes an additional articulatory distance that measures the distance of the test configuration from the central fricative configuration estimated in step one. In this way, contextual information in the surrounding voiced regions are used to improve the fricative estimate. As discussed in Section 6.2, this augmented distance measure, when properly weighted, significantly improves static fricative estimates.

6.3.3 Step Three: Reprocessing of Voiced Speech Segments

With an estimate of the central fricative configuration, it is desirable to have the voice segments smoothly transition into and out of the fricative segment. Therefore, the voiced segments are re-estimated by re-applying the dynamic programming procedure with the fricative frames fixed to the result of step two. In a bootstrapping manner, the voiced acoustic-to-articulatory mapping has been used to aid fricative acoustic-to-articulatory mapping which is, in turn, used to re-estimate the voiced

segments. As a result, both the voiced and fricated segments have an opportunity to influence the other.

While only a single fricative configuration is estimated in step two, this configuration should be in force over a significant fraction of the fricative segment. Initial versions of the procedure fixed only the central fricative frame and allowed the voiced acoustic-to-articulatory mapping algorithm to determine all others. This produced very slow transitions into and out of the fricative which drastically altered the way it was perceived. We observed that that the dynamics of the rapid transition from vowel to fricative offer substantial cues to the identity of the fricatives. Incorrect dynamics can alter the perception of the fricative, even when the fricative acoustics have been well matched. This observation is supported by the perceptual experiments of Harris [55].

To prevent the slow transitions and unintentional perceptual effects, the estimated central fricative configuration is duplicated in the frames representing a centered fraction of the entire fricative segment. The remaining outer frames of the fricative segment are allowed to be selected by the voiced acoustic-to-articulatory mapping procedure. The fraction of the fricative segment that is fixed is set to 70% by default, but often needs to be hand adjusted to as large as 95% to improve synthesis.

6.3.4 Step Four: Constriction Area Selection

Constriction area is a feature of significance to both the vocal-tract and glottal-source simulations. The vocal-tract transfer function is very sensitive to the small constriction areas associated with fricative configurations. Similarly, frication amplitude is sensitive to constriction area which affects the resistance to flow at the

constriction. This dual dependency on constriction area suggests the need for simultaneous source and tract optimization in acoustic-to-articulatory mapping schemes.

Source-tract interaction is a difficult issue that can complicate the formulation of acoustic-to-articulatory mapping schemes. To reduce complexity in acoustic-to-articulatory mapping, it is desirable to avoid simultaneous optimization of source and tract variables. In the synthesizer, the fricative source and the vocal-tract transfer function are both a function of constriction area. Therefore, by assuming a fixed value for constriction area in the central fricative frame, the source and tract can be decoupled. This approach allows us to focus on vocal-tract optimizations without the difficulties of the source simulation, at the expense of limiting somewhat the amount of source optimization that can be performed.

The fixed constriction area used, $0.125\,cm^2$, is consistent with reported measurements of constriction area in real fricatives [47, 48, 33]. The vocal-tract optimization is constrained in step two to estimate an articulatory trajectory that passes through this point by using a fricative linked-codebook constrained to have constriction area between 0.10 and $0.15\,cm^2$.

To avoid vowel configurations that produce audible frication, the configurations in the vowel codebook used in steps one and three are limited to cross-sectional areas greater than $0.3\,cm^2$. Since the fricative codebook in step two is limited to configurations with constriction areas less than $0.15\,cm^2$, the overall lookup procedures of steps one through three cannot offer alternatives with constrictions in the range of 0.15–$0.3\,cm^2$ for the vowel-fricative transitions. For smooth onset of frication, constriction area should smoothly change from greater than $0.3\,cm^2$ in the voiced segments to less than 0.15 during the fricated segments.

To affect this smooth area transition, constriction area is defined in the transition region bordering vowel and fricative. As shown in Figure 6.6, sigmoidal interpolation functions are used to define a smooth constriction area transition into and out of the fricative. If $A_c^{vowel}(t)$ is the constriction area trajectory estimated in steps one through three, and A_c^{fric} is the constriction area of the central fricative frame, the desired constriction area trajectory, $A_c^{new}(t)$, is defined as the combination of $A_c^{vowel}(t)$ and A_c^{fric} using sigmoid functions for interpolation as follows

$$A_c^{new}(t) = (A_c^{vowel}(t) - A_c^{fric})(g_{vf}(t) + g_{fv}(t)) + A_c^{fric} \tag{6.2}$$

where

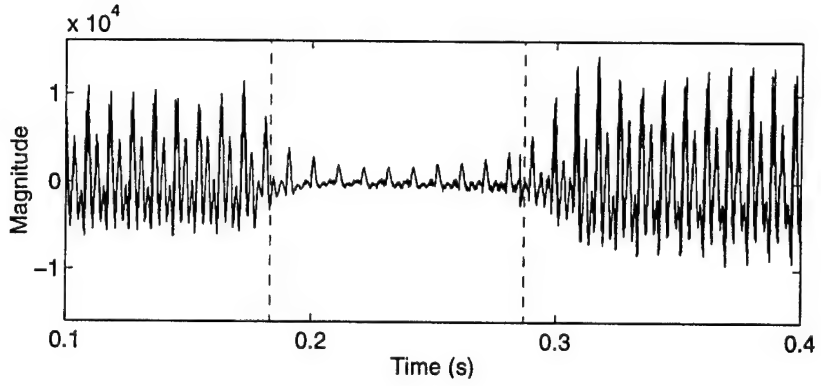
$$g_{vf}(t) = \frac{1}{1 + \exp^{\eta_{vf}(t-\tau_{vf})}} \tag{6.3}$$

and

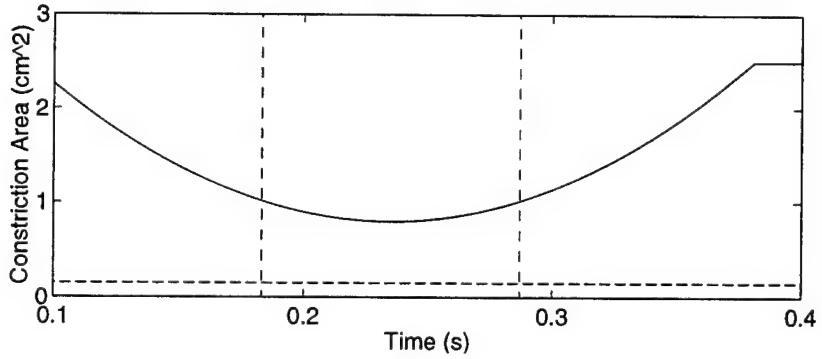
$$g_{fv}(t) = \frac{1}{1 + \exp^{-\eta_{fv}(t-\tau_{fv})}}. \tag{6.4}$$

The sigmoids $g_{vf}(t)$ and $g_{fv}(t)$ perform interpolation from vowel to fricative and fricative to vowel respectively. The location of the transition and the rate of transition is controlled by η and τ terms of each sigmoid. Default values of are η and τ based on the fricative voicing class and segmentation boundaries have been chosen.

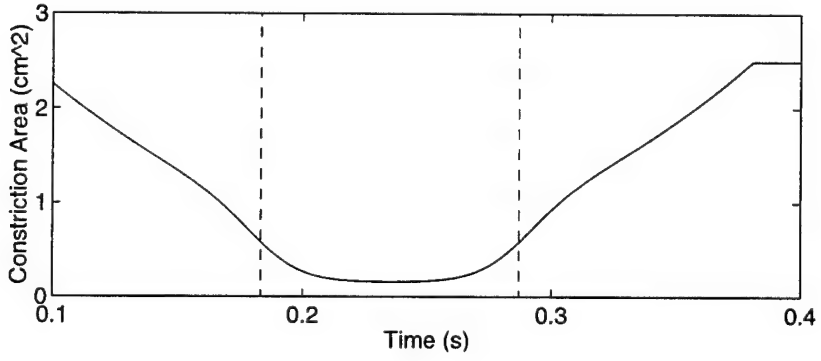
By specifying constriction area in this way, we are assuming that the majority of articulator motion has already occurred and that the transition to a narrow constriction for frication occurs without significant additional articulator motion. The reapplication of dynamic programming in step three with the central fricative frame fixed should produce a trajectory with a constriction area that decreases as the fricative is approached, and increases as the fricative is completed. Therefore, this assumption is not unreasonable.



(a) Waveform for /ivi/ with vowel and fricative sections delineated.



(b) Constriction areas, $A_c^{vowel}(t)$ (solid) and A_c^{fric} (dashed), estimated in steps one through three.



(c) Constriction area, $A_c^{new}(t)$, after sigmoidal interpolation.

Figure 6.6: Specification of constriction area transition into and out of fricative using sigmoidal interpolation.

Since constriction area is not an explicit variable in the LAM, the desired constriction area trajectory is imposed by a numerical optimization procedure that minimizes the distance between the actual configuration area and the desired area on each frame.

6.3.5 Step Five: Glottal Source Parameter Selection (Resynthesis)

The result of step four is an articulatory trajectory that represents an acoustic--to-articulatory mapping solution. While not part of the actual acoustic-to-articulatory mapping procedure, the production of synthetic speech from our estimated articulatory trajectory is one method for verifying the suitability of the acoustic-to-articulatory mapping result. Since this procedure does not optimize source parameters to match aspects of the original utterance such as frication amplitude or degree of voicing, source parameters must be generated in some other consistent manner.

The constriction area trajectory has been determined in step four. Therefore, only the source parameters — glottal rest area, A_{g0}, glottal tension, q, and lung pressure, P_s — may be varied to produce voicing of the correct pitch and amplitude, natural sounding transitions into and out of fricative segments, and the correct perception of voiced or unvoiced frication.

A simplified approach to source parameter selection is used. Given the type of each fricative segment, either voiced or unvoiced, settings for source parameters in the central fricative frame are set to predefined values. Sigmoidal interpolation is used to provide smooth transitions between vowel and fricative segments. This prespecification of source parameters is employed to achieve acceptable synthesis quality with the least amount of automation complexity. This approach avoids many of

the complications of source optimization such as controlling voicing and quantifying the relative contributions of voicing and frication. This approach also avoids known deficiencies in the source model. For example, we found it difficult to produce voiced fricatives with significantly reduced amplitudes relative to voiced segments as observed in our dataset. We suspect this might be due to limitations in the glottal-source model and the constriction resistance estimation. Such limitations would be difficult to overcome solely using optimization.

Fundamental frequency during voicing is controlled by the glottal tension factor, q. Using a pitch extraction routine and one-dimensional optimization, glottal tension is adjusted to match the pitch of the voiced portions of the original utterance. For unvoiced fricatives, a default glottal tension of 1.6 is specified for the central fricative frame. This value is high enough to damp glottal vibration when glottal rest area is simultaneously increased. Transitions of q between the central unvoiced fricative frame and nearby voiced frames is achieved using sigmoidal interpolation, as used for constriction area interpolation. The gain factors on the sigmoidal interpolation is predefined depending on the fricative voicing classification.

Lung pressure is optimized to best match the power in each voiced frame without changing too quickly. Lung pressure is assumed constant during VFV transitions and is not allowed to contribute to differences in amplitude between vowels and fricatives which change too quickly to be attributed to changes in lung pressure.

Glottal rest area, A_{g0}, is the primary control of voicing during frication. During voiced sounds, including voiced fricatives, glottal rest area is set to $0.03\,\mathrm{cm}^2$. For unvoiced fricatives, glottal rest area in the central fricative frame is set to $0.14\,\mathrm{cm}^2$

which prevents glottal vibration when glottal tension is simultaneously increase. Transitions of A_{g0} between the central unvoiced fricative frame and nearby voiced frames is achieved using sigmoidal interpolation.

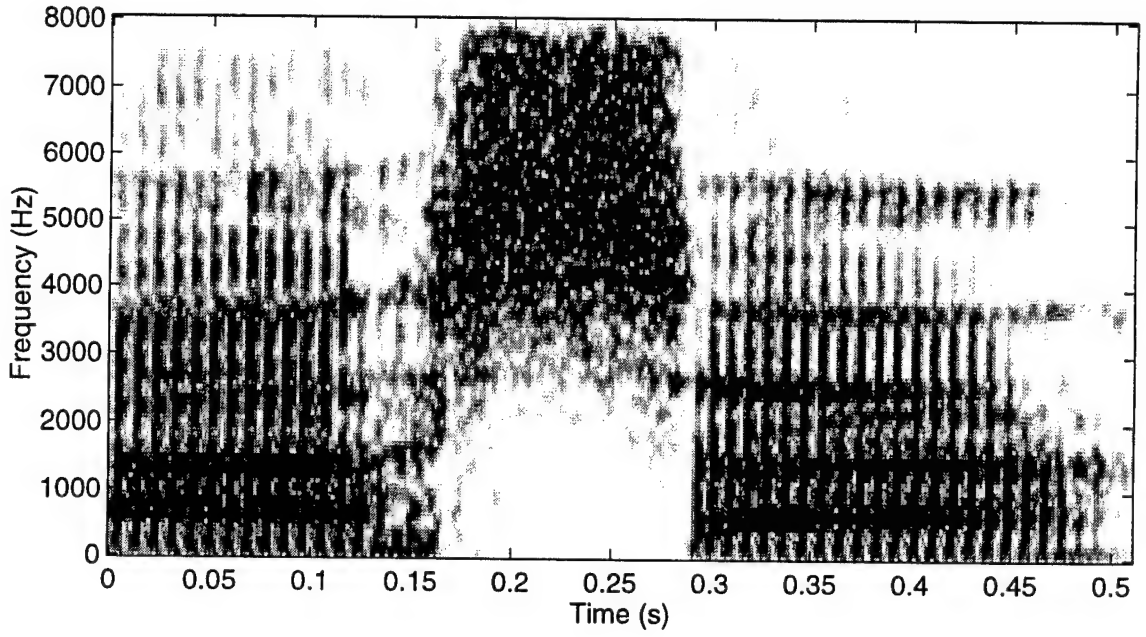
6.4 Example

To demonstrate the operation of the five step procedure, the acoustic-to-articulatory mapping of /ɑsɑ/ spoken by MJC will be detailed. The results of each stage will be illustrated and discussed.

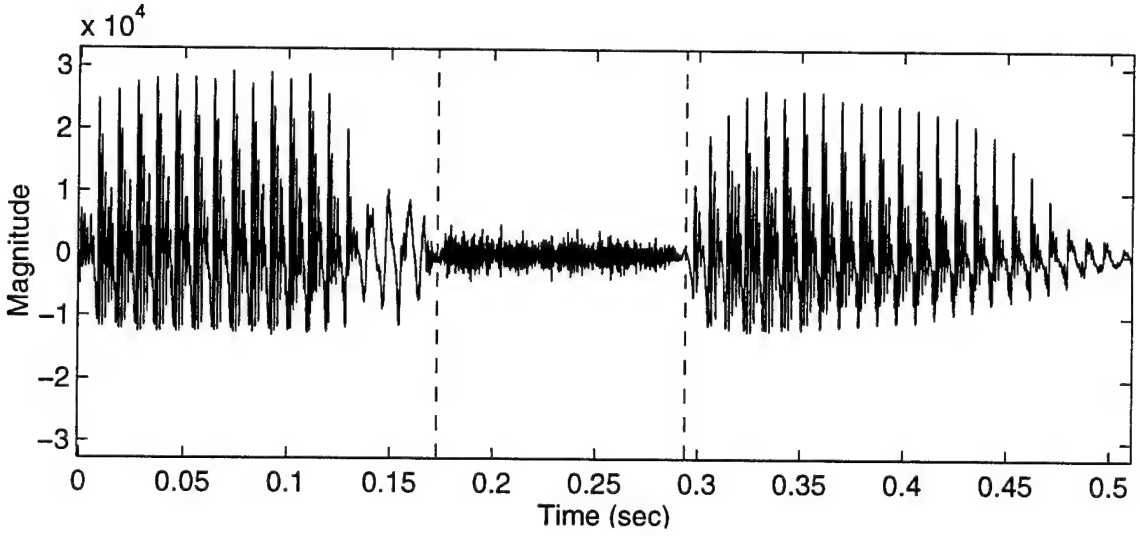
Figure 6.7 shows the spectrogram and waveform of /ɑsɑ/. The utterance has been divided by hand into voiced and fricated regions. /s/ is an unvoiced fricative and appears to be completely devoiced in the central fricative frame. Weighted FFT cepstral coefficients were calculated over 32 ms frames with 16 ms overlap. A total of 31 frames of features were generated.

Voiced acoustic-to-articulatory mapping using weighted cepstral coefficients and continuity in resonance was performed on the utterance. Figure 6.8 shows the estimated resonances (formants) and the articulatory configurations in the central frame of each phone. The resonances appear correctly located in the voiced segments and move significantly in the fricative segment. The articulatory configuration for the central fricative frame cannot produce frication but does show movement from the low, back tongue position of /ɑ/ toward the alveolar constriction of /s/.

A FWCEP feature was calculated on a 64 ms frame located in the center of the fricative segment. Fricative linked-codebook lookup was performed with this feature, using LAM distance from the voiced estimate for contextual information. Figure 6.9 shows the estimated articulatory configuration and acoustic fit. The estimated con-

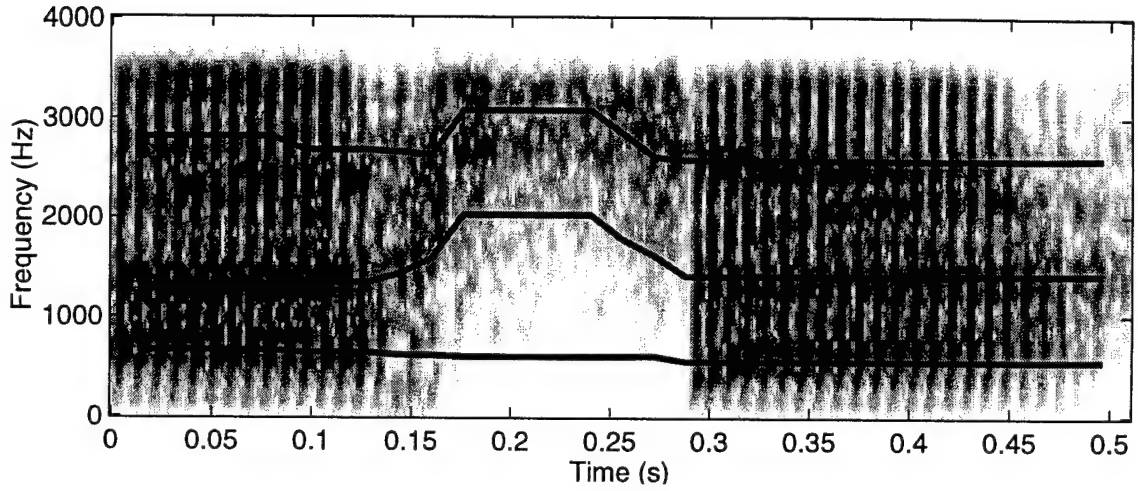


(a) Spectrogram

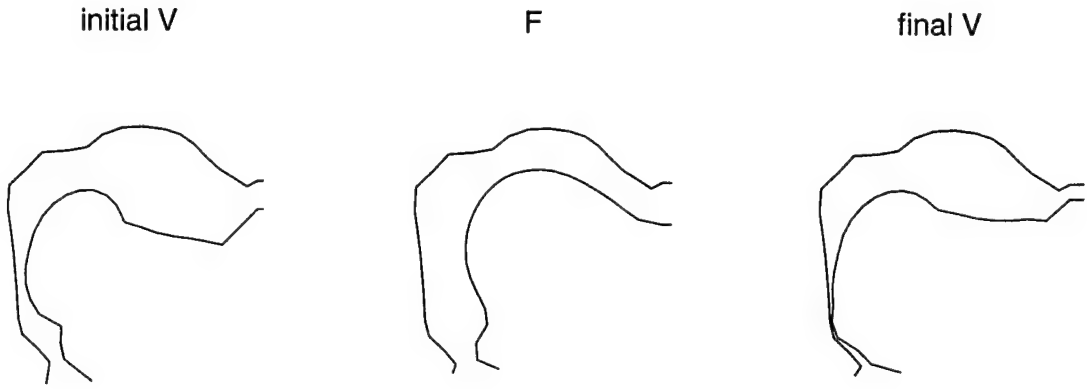


(b) Waveform

Figure 6.7: Spectrogram and waveform of /ɑsɑ/ spoken by MJC.



(a) Spectrogram and estimated resonances.



(b) Central articulatory configuration of each segment.

Figure 6.8: Results of applying the voiced, dynamic acoustic-to-articulatory mapping algorithm to /ɑsɑ/.

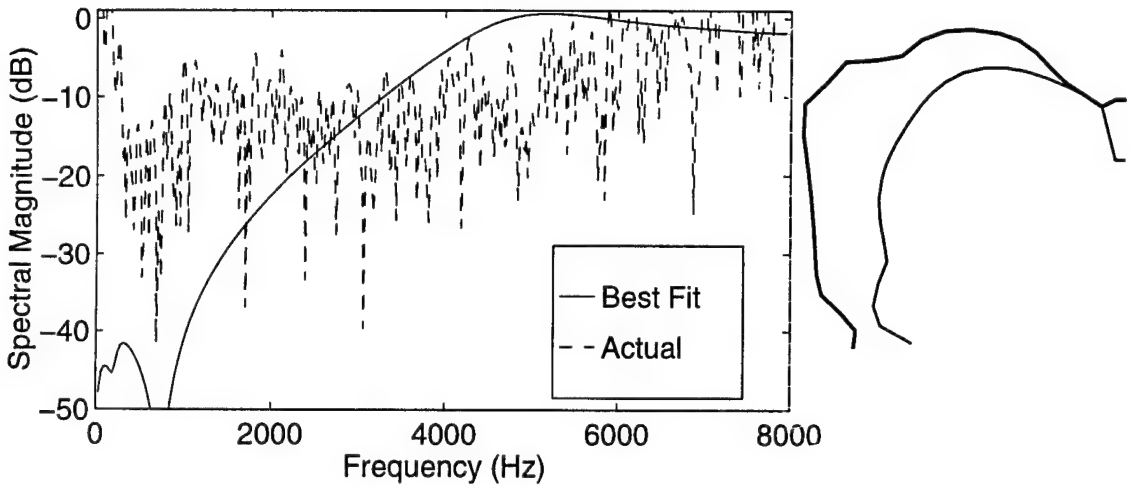


Figure 6.9: Acoustic-to-articulatory mapping result on central fricative frame of /asa/.

figuration has a correctly placed alveolar constriction and the transfer function has a large magnitude at high frequencies as expected for an alveolar fricative. The zero in the transfer function near 800 Hz is typical for configurations with the frication source located distant from the constriction (0.79 cm from the lips in this case). While a good overall solution, it is remarkable that this transfer function, with its large dynamic range, was found to be the best match to the real spectrum.

Dynamic programming was re-applied to the lookup results of step one, except that the articulatory configuration in the central 70% of the fricative segment was fixed to the result of step two. Figure 6.10 shows the estimated resonances after this step. Four frames before the central fricative and two frames after have been changed from the results of step one in Figure 6.8(a).

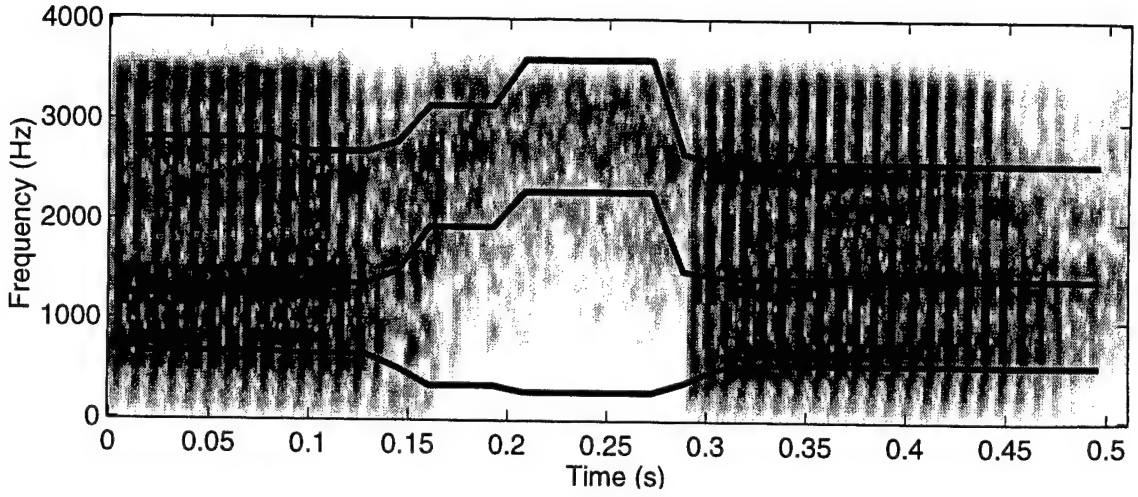


Figure 6.10: Resonance trajectories after step three.

The default central fricative frame constriction area is $0.125\,\mathrm{cm}^2$. The transition from this constriction area to the natural constriction area of the voiced segments is specified through sigmoidal interpolation. The rate and location are set to default values, which were selected by hand for a similar VFV token. Optimization adjusts the configurations resulting from step three to have the desired constriction area trajectory. Figure 6.11 shows the constriction area trajectory before and after constriction area optimization.

Figure 6.12 illustrates the articulatory trajectory resulting from step four. This is the acoustic-to-articulatory mapping result. Despite the dynamic programming and use of contextual information, the transitions from /a/ to /s/ and back appear quite abrupt. The configurations labeled one through three correspond to the last half of the initial vowel. The configurations labeled four through six are fricative formations. The last three configurations correspond the the transition from fricative to final vowel. Frames of 32 ms, while adequate for vowels and glides, may not be

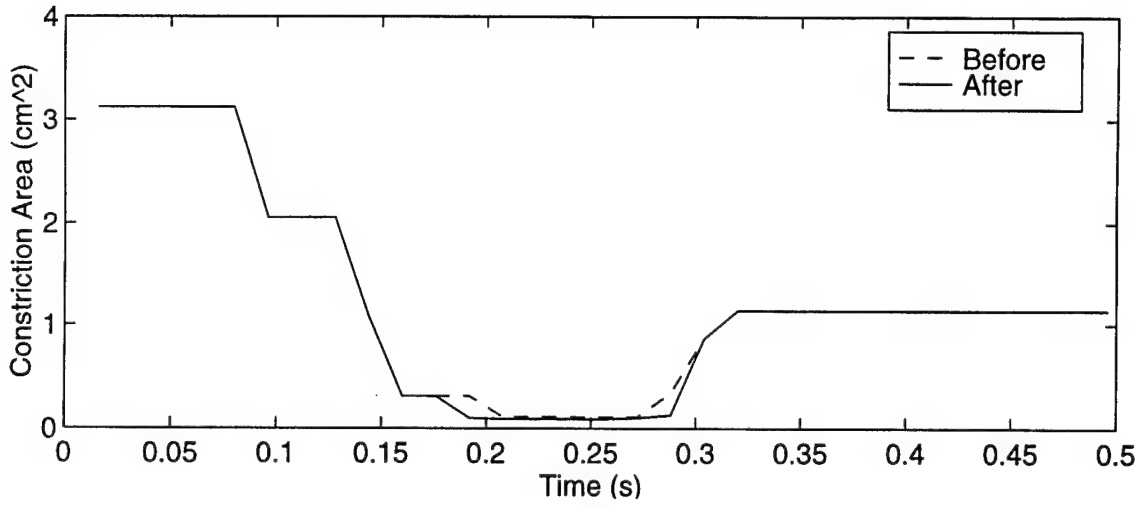


Figure 6.11: Constriction area before and after step four processing on /ɑsɑ/.

small enough to reproduce the necessary changes in transition regions between vowels and fricatives. Additional optimization on this dynamic programming result should improve the smoothness of the sequence as well.

To generate synthetic speech and validate our solution, glottal-source parameters in the fricative segment and transition regions were set to predefined settings for an unvoiced fricative. Values of glottal tension, q, and lung pressure, P_s, during the voiced segments were adjusted to match fundamental frequency and amplitude respectively. The source parameters used are illustrated in Figure 6.13.

Figure 6.14 illustrates the waveform and spectrogram of the resynthesized speech. There is a strong similarity between the spectrograms of Figure 6.14(a) and Figure 6.7(a). The utterance is clearly perceived as /ɑsɑ/. Due to the lack of fricative source optimization in the transition region, the fricative durations are different. The lack of optimization after dynamic programming is obvious in the abrupt changes in

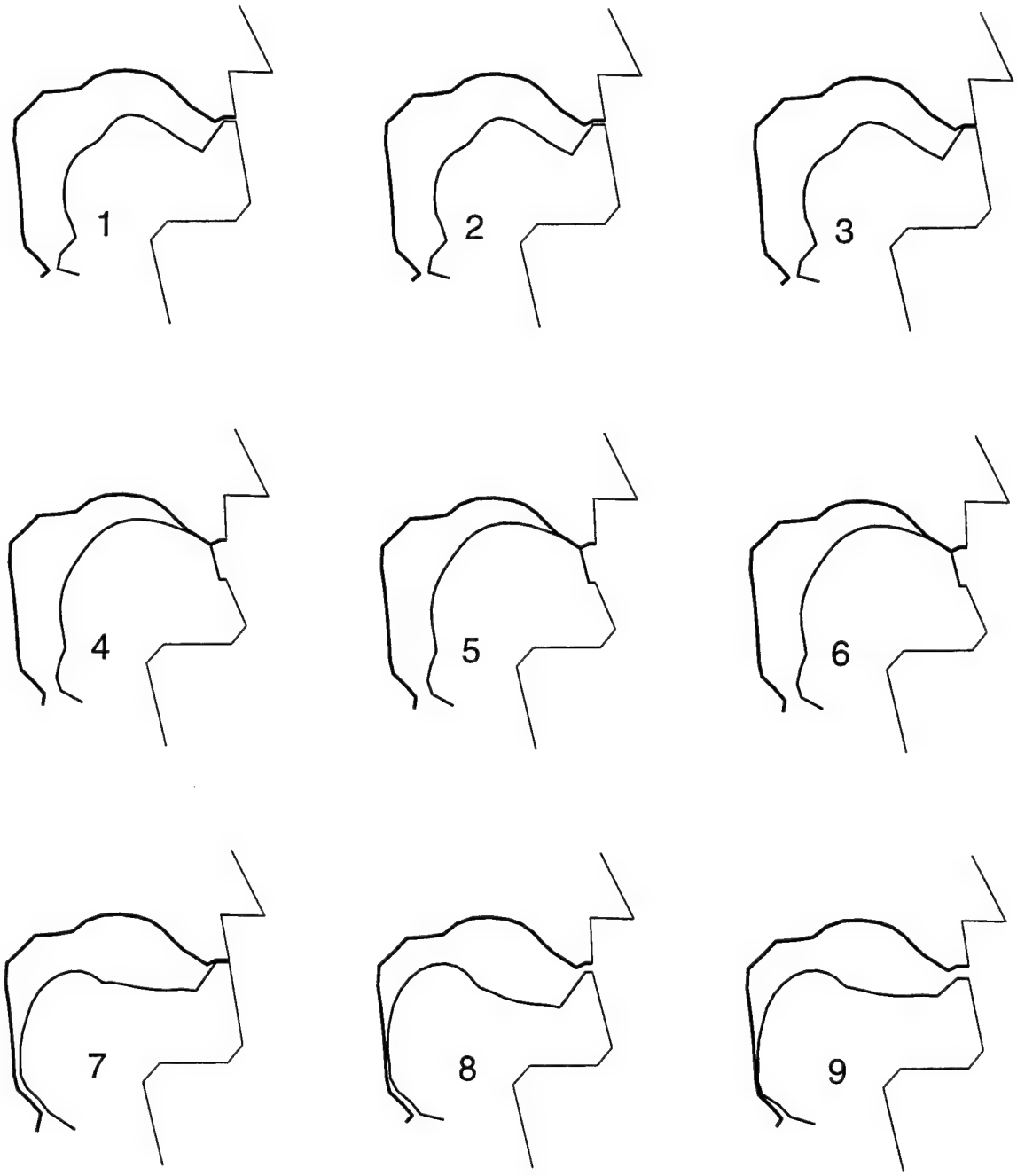


Figure 6.12: Sequence of estimated articulatory configurations from central nine frames of /ɑsɑ/.

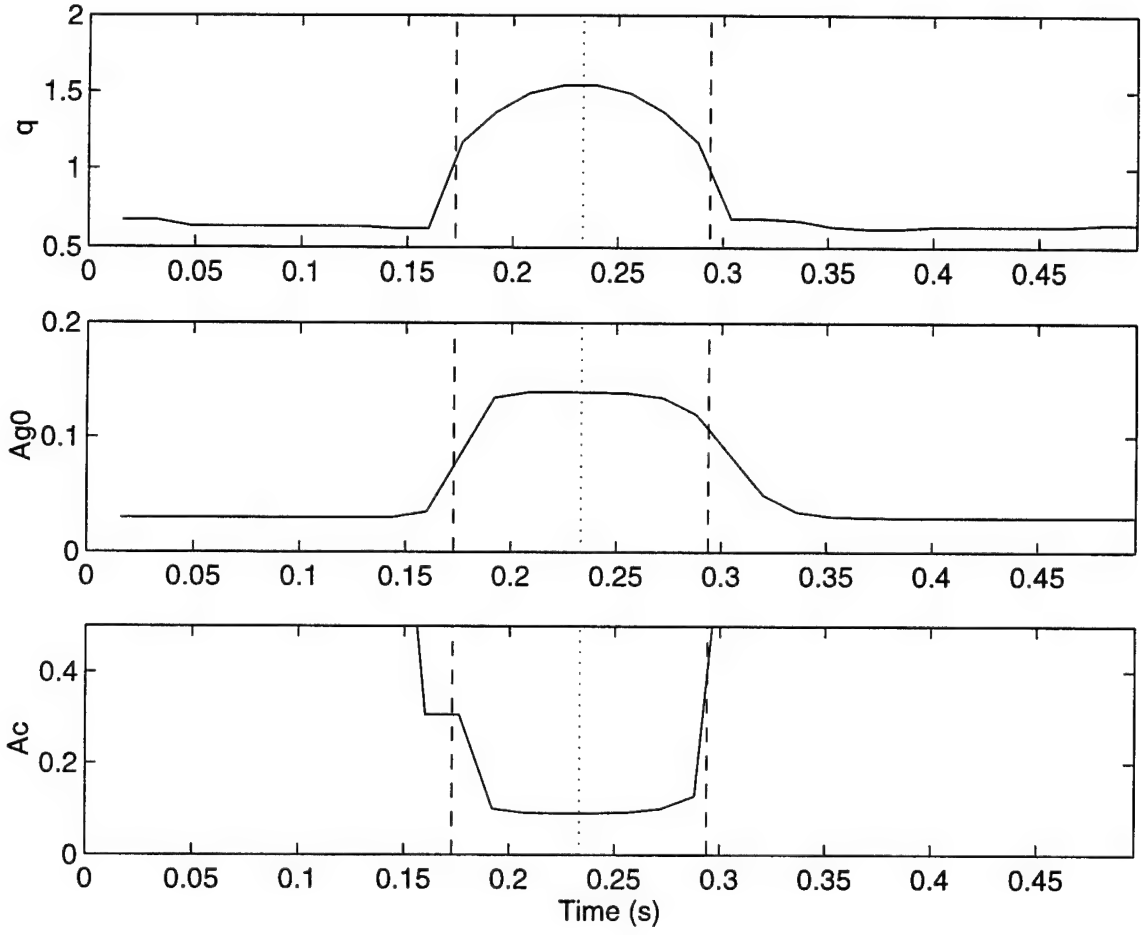
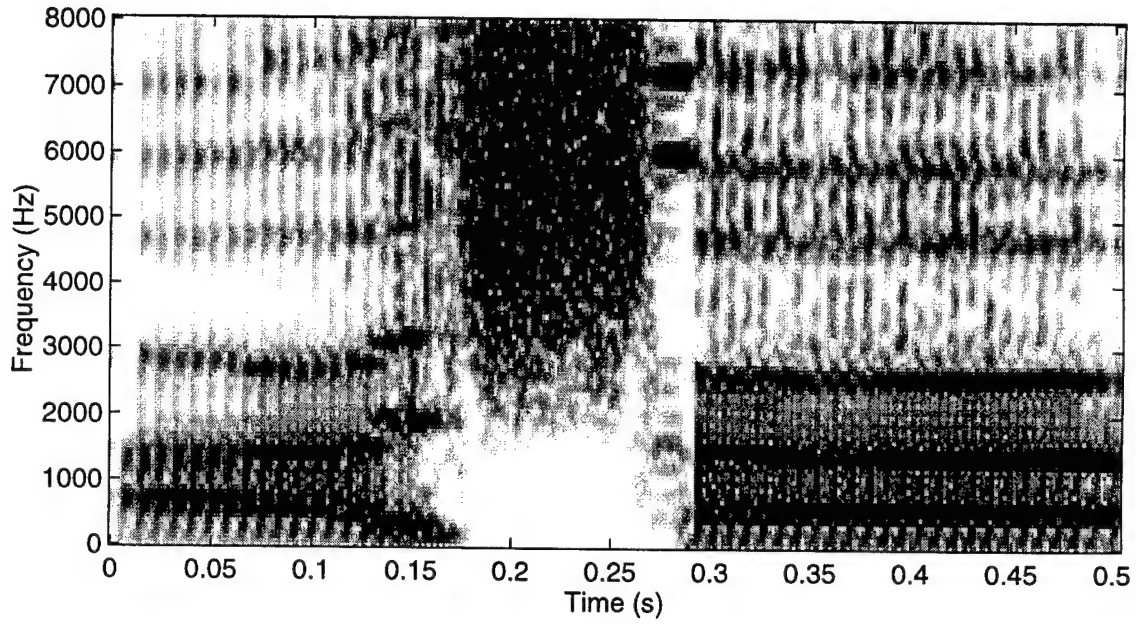
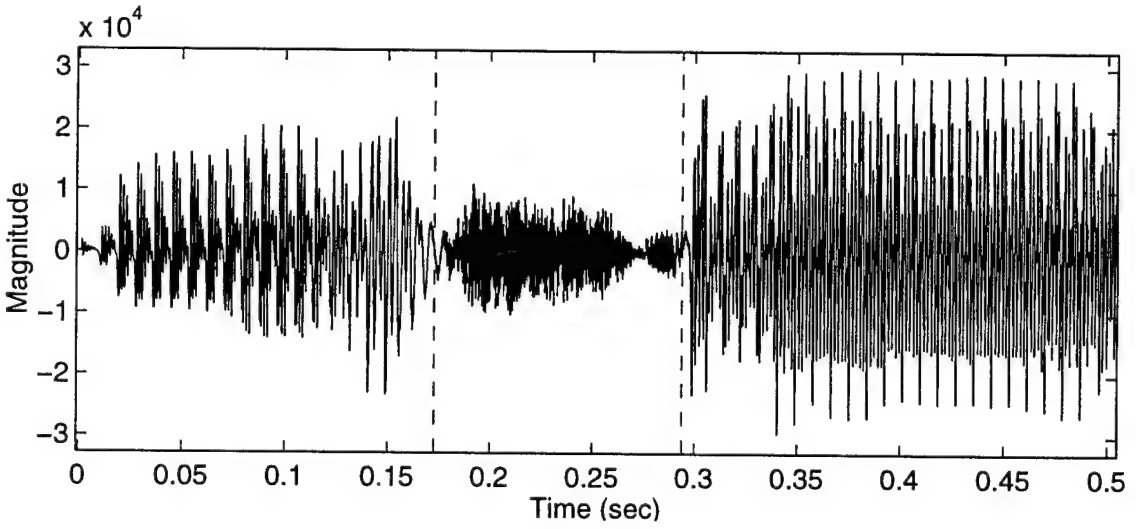


Figure 6.13: Glottal and fricative source parameters used for resynthesis of /ɑsɑ/: glottal tension, q, glottal rest area, A_{g0}, and constriction area, A_c.



(a) Spectrogram



(b) Waveform

Figure 6.14: Spectrogram and waveform of resynthesized /ɑsɑ/.

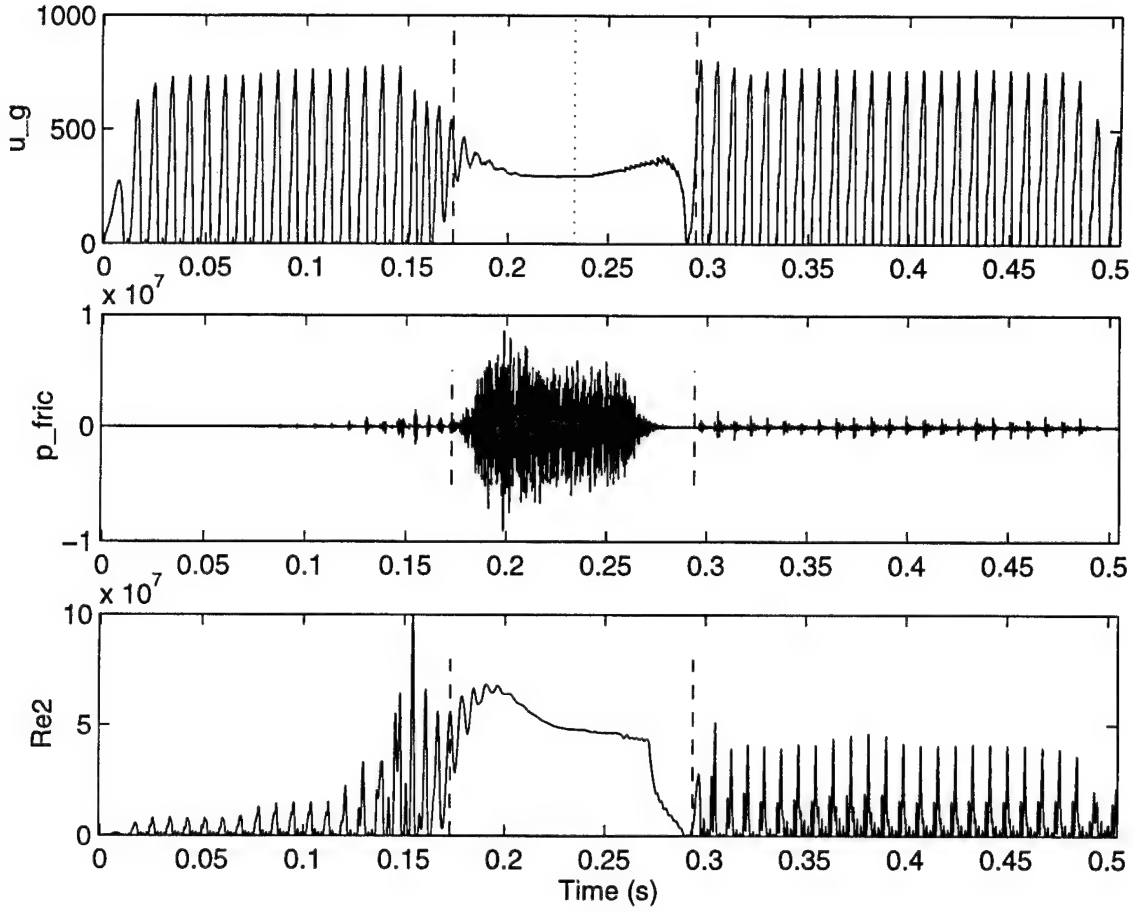


Figure 6.15: Glottal and fricative source simulation for /ɑsɑ/: glottal flow, u_g, fricative output, p_{fric}, and squared Reynolds number at the constriction, Re^2.

waveform shape near 0.13 s and 0.33 s. Artifacts can also be seen in the transition regions of the spectrogram.

In Figure 6.14(b), an interruption in frication can be observed at 0.257 s. This is the point where the constriction begins to open. Controlling the source simulation to maintain frication, yet start oscillation for the following vowel has proven to be a difficult task. Figure 6.15 plots the relative source simulation parameters and how they change during the synthesis of /ɑsɑ/. Widening the constriction is necessary to

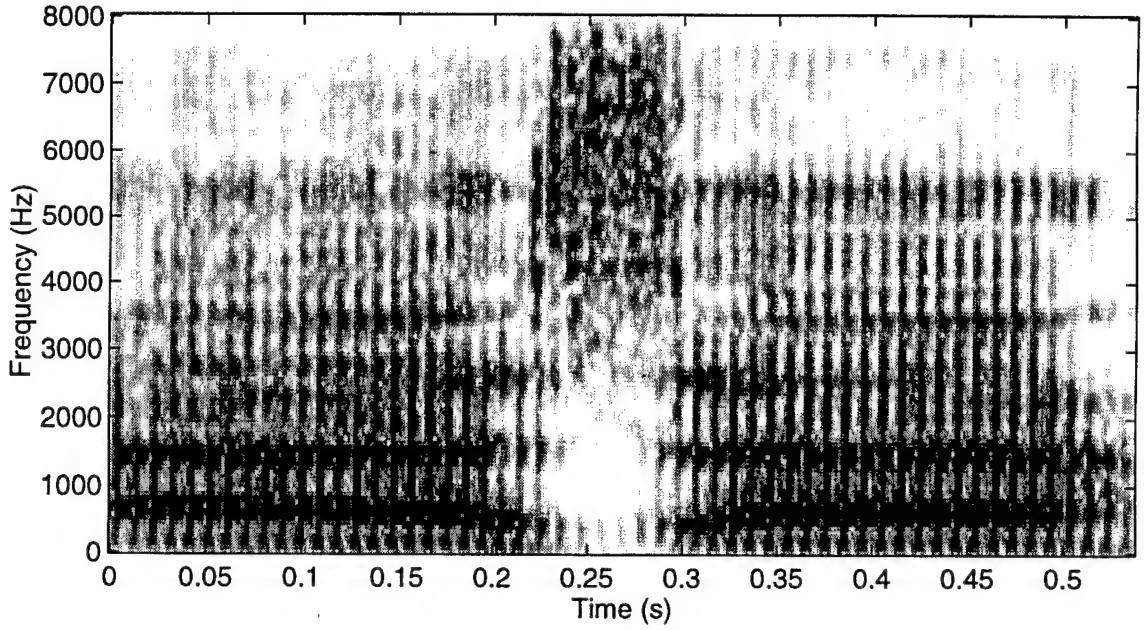
increase airflow and start oscillation, yet the widening drastically reduces Reynolds number and, in turn, frication amplitude. Better control of these source parameters is needed.

6.5 Evaluation

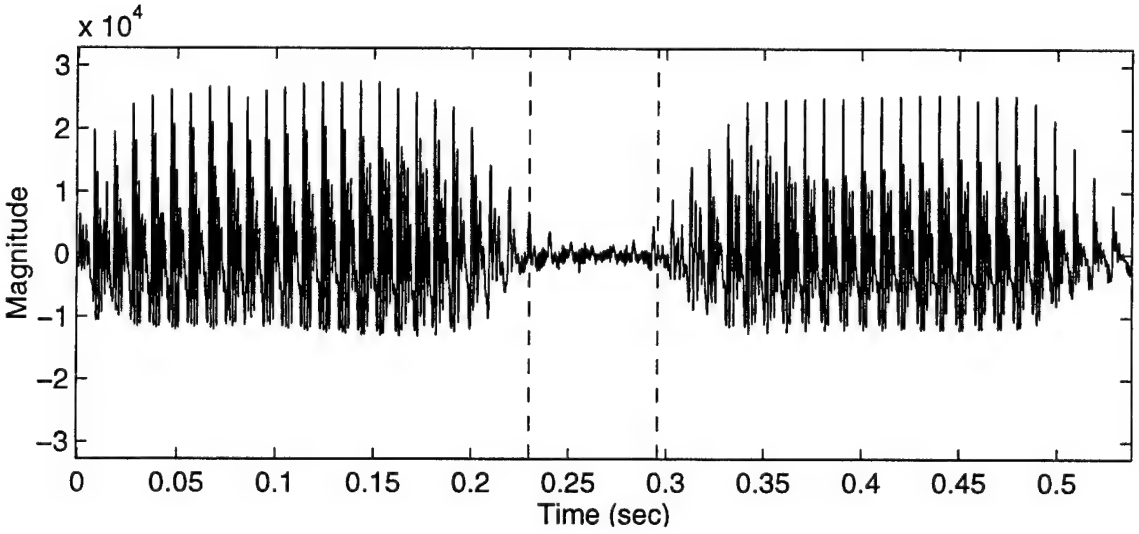
The previous example demonstrated the successful acoustic-to-articulatory mapping and resynthesis of an utterance containing an unvoiced fricative. Figures 6.16 and 6.17 show the original and resynthesized speech respectively for the utterance /ɑzɑ/. The resynthesized speech is clear perceived as /ɑzɑ/. The transitions regions are not as discontinuous as for the unvoiced case, but the resynthesized version exhibits the same abrupt waveform changes. The presence of voicing appears to improve configuration estimates in the transition regions. Frication amplitude, relative to vowel amplitudes, is not well controlled.

In order to get a more quantitative assessment of how the algorithm performs, the five step procedure was applied to all 48 VFV tokens from speaker MJC. Of all tokens, only four produced fricatives with constriction locations that indicated incorrect sibilant classification. This mistake causes the step three processing to incorrectly modify the voiced frames. Since the voiced transitions contain many cues to the identity of the fricative, this type of error can be catastrophic.

Of the estimated vowels, four configurations for /u/ were incorrectly formed as high front vowels. This occurred by mistaking F4 for F3 and F3 for F2. For the VFV cases, the vowel frames are short and have little movement with which to let continuity constraints eliminate unlikely trajectories. In longer utterances, this might not be such a problem. Occasionally, /ɑ/ and /u/ sound rough due to poorly smoothed

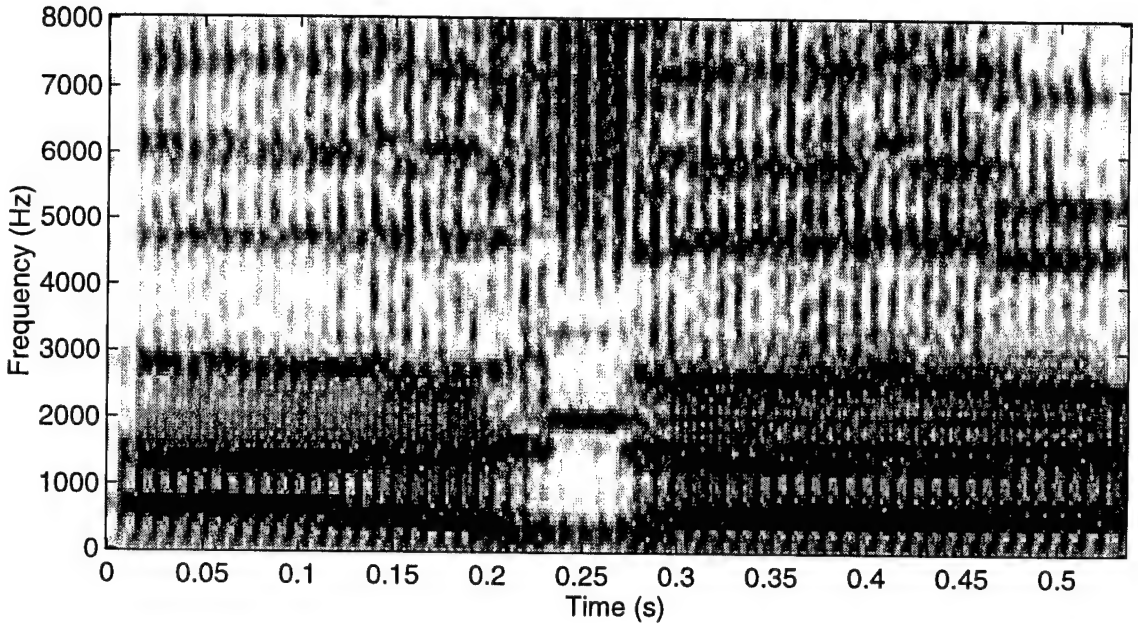


(a) Spectrogram

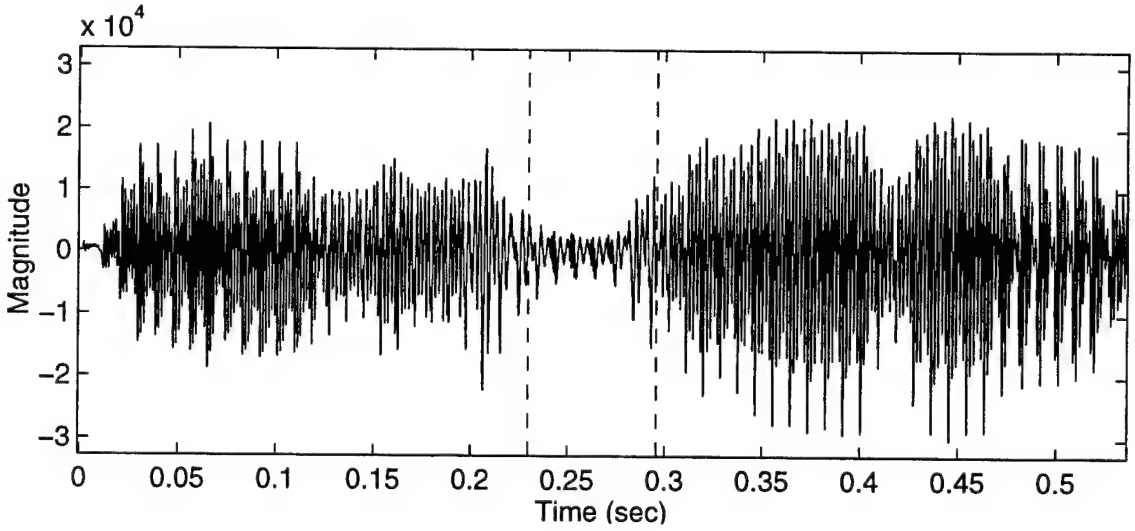


(b) Waveform

Figure 6.16: Spectrogram and waveform of /ɑzɑ/ spoken by MJC.



(a) Spectrogram



(b) Waveform

Figure 6.17: Spectrogram and waveform of resynthesized /ɑzɑ/.

trajectories. This may be due to the fact that /ɑ/ and /u/ must move more than /i/ to reach fricative configurations. Eleven of the 48 tokens produced unstable synthesis. This problem is more likely due to problems in the source simulation that errors in the acoustic-to-articulatory mapping result. The two-mass model is sensitive to parameter settings and easily goes unstable when sub-glottal pressure gets high.

An informal listening test with two listeners resulted in 57% of the (stable) VFV tokens identified correctly. Of the misidentifications, 37.5% were due to confusion between dental and labio-dental fricatives, and 37.5% were due to incorrectly classifying palato-alveolar sounds. Problems with the identification of palato-alveolar sounds is to be expected since Chapter 4 found palato-alveolar fricatives to be difficult to estimate well. The remaining 25% of classification errors were random errors in place of articulation. No errors were made in voicing identification.

If linked-codebook lookup performed properly for both voiced and fricative segments, the algorithm produces a reasonable articulatory trajectory. During resynthesis though, a new set of problems emerged. It is at this stage that slight errors in the articulatory result during transition frames become apparent. For many fricatives, especially the dental and labio-dental fricatives, errors in the transition phase caused the fricative to be perceived incorrectly. Often, errors caused the fricative to sound like an alveolar fricative — even if the fricative spectra does not look alveolar. These transition errors are due to errors in the source parameters as well as the articulatory trajectory.

After step three and step four, further iterative optimization could significantly improve results. Optimization would improve acoustic fit slightly, but more significantly,

would make the articulatory transitions smoother. This will result in waveforms that do not abruptly change in appearance and sound more natural.

There remain some unsolved issues regarding optimization during the transitions. Treating the transitions as voiced, as is done now, is not ideal since they contain contributions from both glottal and frication sources. Using shorter frames or some type of parametric estimation might improve current results under the voiced assumption. Jointly optimizing both the voiced and fricative contributions to the transition would be best, but effective distance measures and measures of relative contribution are needed. This type of joint optimization also would require simultaneous source optimization. The source-tract interaction issues and lack of effective frication measures make this a problem that would require a significant amount of work.

CHAPTER 7

CONCLUSIONS

Acoustic-to-articulatory mapping is a challenging problem in which a time-varying vocal-tract shape is estimated from only the speech waveform. Most progress in acoustic-to-articulatory mapping has been achieved for voiced speech, without obstruent sounds such as stops and fricatives. This dissertation considers the acoustic-to-articulatory mapping problem with a focus on fricatives. Fricative production and perception offer unique challenges to acoustic-to-articulatory mapping which stretch many of the assumptions, measures, and heuristics of existing acoustic-to-articulatory mapping algorithms. This work identifies these issues in fricative acoustic-to-articulatory mapping and extends existing techniques for voiced speech to unvoiced and voiced fricatives in isolation and in continuous speech.

An articulatory speech synthesizer was constructed for analysis-by-synthesis-based acoustic-to-articulatory mapping. The synthesizer is based on the hybrid time-frequency domain articulatory speech synthesizer of [8] and uses the two-mass glottal model of Ishizaka and Flanagan [44] for the glottal-source simulation. Fricative production was modeled using a single noise pressure source that could be located anywhere forward of the constriction. A technique was developed for simultaneously approximating constriction resistance and constriction flow in the aerodynamic simulation that enables

frication amplitude to be controlled in a natural way. A linear articulatory model (LAM) [42], with the addition of one parameter to specify frication pressure source location, was used as front end to the synthesizer for both voiced and fricated speech.

Linked-codebook procedures were used to perform acoustic-to-articulatory mapping experiments on vowels and fricatives in isolated and continuous speech. Linked-codebooks are a table-based procedure that provide a coarse representation of the acoustic-to-articulatory transformation and can be used to provide initial estimates for further optimization. They have been used herein to both study the fricative acoustic-to-articulatory transformation, and design acoustic-to-articulatory mapping algorithms for voiced and fricated speech.

Linked-codebooks were first used to examine the acoustic-to-articulatory mapping of voiced and unvoiced static fricatives. While fricative inverse mapping has been described in previous works, the capabilities and limitations of a fricative acoustic-to-articulatory mapping algorithm have not been studied closely. Our investigation attempted to provide insight on performance by examining the properties of inverse mapping over many cases. Acoustic-to-articulatory mapping performance was evaluated by analyzing articulatory estimation error for a large number of synthetic fricatives and phonetic class clustering for a collection of real fricatives. The quality of inverse mapping results for individual cases was measured based on acoustic fitness.

Due to model-mismatch and non-uniqueness in the fricative acoustic-to-articulatory transformation, reasonable results for static fricatives were not always achieved. Histograms of articulatory distance showed that the amount of articulatory error from linked-codebook lookup can vary widely and is dependent on the acoustic feature employed. The presence of multiple clusters for a single phonetic class in scatter plots of

estimated constriction and frication source location demonstrated that acoustic-to--articulatory mapping can make consistent errors. The scatter plots also illustrated occasional large errors that were physically implausible. With some modifications to the algorithm, acoustic-to-articulatory mapping results were improved to an acceptable level. By enforcing adequate articulatory constraints, many of the consistent errors demonstrated in the scatter plots were avoided. The proper choice of acoustic distance was found to significantly impact performance as well. Both features were modified by removing the contribution of energy from frequencies below 1 kHz. The modified features demonstrated a reduced sensitivity to the presence of voicing energy in real voiced fricatives, with equivalent performance for unvoiced fricatives.

The linked-codebook procedure for static acoustic-to-articulatory mapping was extended to continuous, voiced speech using a dynamic programming procedure [13]. This procedure minimizes a cost function that combines a measure of distance between actual and synthetic acoustic features within frames and a measure of continuity in articulatory features between frames. Successful acoustic-to-articulatory mapping was achieved using the linked-codebook/dynamic programming procedure with formant frequencies as acoustic features and LAM parameters as articulatory features. Since formant frequencies cannot be extracted from many non-voiced speech sounds, some alternative acoustic features were considered. The weighted-cepstral feature of Meyer et al. [30], which was specifically designed for acoustic-to-articulatory mapping, performed well, but with LAM parameters as articulatory features, was more likely to make errors in estimated formant frequencies. These types of errors are considered severe since they alter the intelligibility of the resynthesis. Performance was improved by using formant frequencies from the synthetic transfer functions as an *articulatory*

feature. These formant frequencies can be calculated for any type of speech or articulatory configuration and can prevent many unnatural frame transitions. The use of formant frequencies as articulatory features in the continuity measure appears to have the same effect as the processing contained in our formant extraction algorithm and gives performance similar to our initial algorithm using formant frequency as acoustic features.

Direct application of the voiced acoustic-to-articulatory mapping algorithm to speech containing intervocalic fricatives cannot correctly process the entire utterance. Acoustic-to-articulatory mapping in the voiced segments was unaffected by the presence of fricatives in the utterance, except for a few utterance containing palato-alveolar fricatives. As expected, since no frication information was used in the inverse mapping, the configurations estimated during the fricative segments were not reasonable; however, the linked-codebook/dynamic programming procedure was able to follow formant transitions into and out of fricatives resulting in estimated trajectories with constrictions that narrow during the fricative segments. This suggests that voiced acoustic-to-articulatory mapping over fricative segments may provide contextual information that can be used to aid fricative acoustic-to-articulatory mapping.

To test whether contextual information from the voiced acoustic-to-articulatory mapping algorithm can improve fricative estimates, the static acoustic-to-articulatory mapping of VFV central fricatives was performed with and without the inclusion of an articulatory distance measure between the estimated fricative configurations and its voiced inverse mapping estimate. Performance was measured in terms of the ability of the static acoustic-to-articulatory mapping procedure to correctly classify the fricative in VFV tokens as sibilant or non-sibilant. Without contextual information, static

acoustic-to-articulatory mapping correctly identified sibilance for 85% of the VFV token. With the addition of contextual information, correct identification rose to 94%.

The results of the static fricative acoustic-to-articulatory mapping experiments were used to extend the dynamic acoustic-to-articulatory mapping algorithm to continuous, voiced speech containing intervocalic fricatives. The fricative acoustic features, constraints, and linked-codebooks developed in Chapter 4 were included in the new algorithm. Based on the ability of contextual information to improve inverse mapping results for both fricatives and vowels, a five step procedure was developed for the dynamic acoustic-to-articulatory mapping algorithm to continuous, voiced speech containing intervocalic fricatives. Multiple stages of processing are used to bootstrap articulatory estimates using contextual information. The results of voiced acoustic--to-articulatory mapping are used to assist fricative inverse mapping. The resulting fricative articulatory estimates are then used to improve the acoustic-to-articulatory mapping in the voiced segments. The algorithm requires an utterance that has been segmented into fricated and voiced segments. Estimation of glottal-source parameters for controlling frication amplitude and state of voicing is not fully automated. Given the voicing status of the fricative, a constriction area trajectory is imposed on the result and predefined source parameters are provided. The result is fed to the synthesizer to produce speech that is correctly perceived in terms of voicing and phonetic class, but whose source characteristics are not matched to the original utterance.

Development and testing of the procedure used a collection of vowel-fricative-vowel (VFV) tokens from a single male talker. In most cases, the estimated articulatory trajectories appeared quite natural and produced the correctly located constriction for

fricatives. Occasional errors occurred due to vowel or fricative misidentification early in the optimization process. Transitions between vowels and fricatives were found to be very difficult to reproduce accurately in this frame-based system. Even when articulatory estimates of central fricative and surrounding voiced speech were accurate, errors in the transition between the two had a profound effect on the way the resynthesis was perceived. Good resynthesis, producing speech perceived as the correct fricative class, often required hand tuning of parameters controlling the dynamics of the vowel-fricative and fricative-vowel transitions. The problem was most apparent for the non-sibilant sounds. If parameters were not chosen carefully, the fricative would often sound like an alveolor fricative, rather than the intended fricative.

Dynamic fricative acoustic-to-articulatory mapping has proven to be a difficult problem. Current models of fricative production are limited in their ability to reproduce fricative spectra and speaker dependent differences. These modeling errors, along with known perceptual ambiguities and the finite amount of information in fricative segments, limit acoustic-to-articulatory mapping of fricatives in isolation. By using fricative specific knowledge in the form of constraints, acoustic features and acoustic distance metrics, consistent articulatory estimates can be obtained for static fricatives. The use of contextual information from adjacent voiced segments can also improve fricative estimates in dynamic acoustic-to-articulatory mapping. Much remains to be done to improve fricative acoustic-to-articulatory mapping in both the static and dynamic cases. Improve models of fricative production and articulation will significantly improve all aspects of fricatives acoustic-to-articulatory mapping. Additional work is needed to address the difficult issue of source optimization and improve results in regions of transition between vowels and fricatives.

APPENDIX A

Phonetic Symbols

Symbol	As in	Phonetic class
/s/	rice	unvoiced alveolar fricative
/z/	rise	voiced alveolar fricative
/ʃ/	mission	unvoiced palato-alveolar fricative
/ʒ/	vision	voiced palato-alveolar fricative
/θ/	teeth	unvoiced dental fricative
/ð/	teethe	voiced dental fricative
/f/	strife	unvoiced labio-dental fricative
/v/	strive	voiced labio-dental fricative
/x/	—	unvoiced velar fricative
/χ/	—	unvoiced uvular fricative
/w/	water	labial-velar approximate
/i/	beat	high front vowel
/ɑ/	father	low back vowel
/u/	boot	high back vowel

Table A.1: Phonetic symbols used in this document, along with examples of their usage and their phonetic classification. Note that the phones /x/ and /χ/ are not used in English and, therefore, do not have examples of usage.

BIBLIOGRAPHY

[1] J. Schroeter and M. M. Sondhi, "Speech coding based on physiological models of speech production," in *Advances in Speech Signal Processing* (S. Furui and M. M. Sondhi, eds.), pp. 231–268, New York, NY: Marcel Dekker Inc., 1992.

[2] K. N. Stevens, "Toward a model for speech recognition," *Journal of the Acoustical Society of America*, vol. 32, no. 1, pp. 47–55, 1960.

[3] K. Shirai and T. Kobayashi, "Estimating articulatory motion from speech wave," *Speech Communication*, vol. 5, pp. 159–170, June 1986.

[4] J. Schroeter and M. M. Sondhi, "Techniques for estimating vocal-tract shapes from the speech signal," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 133–150, January 1994.

[5] J. L. Flanagan, *Speech Analysis Synthesis and Perception.* New York: Springer-Verlag, 2 ed., 1972.

[6] J. Liljencrants, *Speech synthesis with a reflection-type line analog.* PhD thesis, Royal Institute of Technology, Stockholm, Sweden, 1985.

[7] E. L. Bocchieri, *An Articulatory Speech Synthesizer.* PhD thesis, The University of Florida, 1983.

[8] M. M. Sondhi and J. Schroeter, "A hybrid time–frequency domain articulatory speech synthesizer," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP–35, pp. 955–967, July 1987.

[9] M. R. Schroeder, "Determination of the geometry of the human vocal tract by acoustic measurements," *Journal of the Acoustical Society of America*, vol. 41, no. 4, pp. 1002–1010, 1967.

[10] P. Mermelstein, "Determination of the vocal-tract shape from measured formant frequencies," *Journal of the Acoustical Society of America*, vol. 41, no. 5, pp. 1283–1294, 1967.

[11] H. Wakita, "Estimation of vocal–tract shapes from acoustical analysis of the speech wave: The state of the art," *IEEE Transactions On Acoustics, Speech, and Signal Processing*, vol. ASSP–27, pp. 281–285, June 1979.

[12] B. S. Atal, J. J. Chang, M. V. Mathews, and J. W. Tukey, "Inversion of articulatory–to–acoustic transformation in the vocal tract by a computer–sorting technique," *Journal of the Acoustical Society of America*, vol. 63, pp. 1535–1555, May 1978.

[13] J. Schroeter, J. N. Larar, and M. M. Sondhi, "Speech parameter estimation using a vocal tract/cord model," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (Dallas, TX), pp. 308–311, April 1987.

[14] S. Parthasarathy and C. H. Coker, "On automatic estimation of articulatory parameters in a text-to-speech system," *Computer Speech and Language*, vol. 6, pp. 37–75, 1992.

[15] S. Gupta and J. Schroeter, "Pitch-synchronous frame-by-frame and segment-based articulatory analysis by synthesis," *Journal of the Acoustical Society of America*, vol. 94, no. 5, pp. 2517–2530, 1993.

[16] R. S. McGowan, "Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests," *Speech Communication*, vol. 14, pp. 19–48, 1994.

[17] V. Sorokin, "Inverse problem for fricatives," *Speech Communication*, vol. 14, pp. 249–262, 1994.

[18] K. Shirai and S. Masaki, "An estimation of the production process for fricative consonants," *Speech Communication*, vol. 2, pp. 111–114, July 1983.

[19] P. Badin and C. Abry, "Articulatory synthesis from x-rays and inversion for an adaptive speech robot," in *Proceedings of the International Conference on Speech and Language Processing*, (Philadelphia, PA), pp. 1125–1128, 1996.

[20] P. Mermelstein, "Articulatory model for the study of speech production," *Journal of the Acoustical Society of America*, vol. 53, no. 4, pp. 1070–1082, 1973.

[21] C. H. Coker, "A model of articulatory dynamics and control," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 452–460, 1976.

[22] R. S. McGowan, "Recovering task dynamics from formant frequency trajectories: Results using computer 'babbling' to form an indexed data base," tech. rep., 127th meeting of the Acoustical Society of America, 1994.

[23] G. M. Fant, *Acoustic Theory of Speech Production.* The Netherlands: Mouton and Co., 1960.

[24] V. Sorokin, "Determination of vocal tract shape for vowels," *Speech Communication*, vol. 11, no. 1, pp. 71–85, 1992.

[25] J. Schroeter and M. M. Sondhi, "Dynamic programming search of articulatory codebooks," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 588–591, 1989.

[26] F. Charpentier, "Determination of the vocal tract shape from the formants by analysis of the articulatory–to–acoustic nonlinearities," *Speech Communication*, vol. 3, pp. 291–308, December 1984.

[27] P. P. L. Prado, E. H. Shiva, and D. G. Childers, "Optimization of acoustic-to-articulatory mapping," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. II–33–II–36, 1992.

[28] J. L. Flanagan, K. Ishizaka, and K. L. Shipley, "Signals models for low bit–rate coding of speech," *Journal of the Acoustical Society of America*, vol. 68, pp. 780–791, September 1980.

[29] S. E. Levinson and C. E. Schmidt, "Adaptive computation of articulatory parameters from the speech signal," *Journal of the Acoustical Society of America*, vol. 74, pp. 1145–1154, October 1983.

[30] P. Meyer, J. Schroeter, and M. M. Sondhi, "Design and evaluation of optimal cepstral lifters for accessing articulatory codebooks," *IEEE Transactions on Signal Processing*, vol. 39, pp. 1493–1502, July 1991.

[31] S. Gupta and J. Schroeter, "Low update rate articulatory analysis/synthesis of speech," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 481–484, 1991.

[32] T. Baer, J. Gore, L. Gracco, and P. Nye, "Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels," *Journal of the Acoustical Society of America*, vol. 90, pp. 799–828, August 1991.

[33] S. Narayanan, A. Alwan, and K. Haker, "Articulatory study of fricative consonants using magnetic resonance imaging," *Journal of the Acoustical Society of America*, vol. 98, no. 3, pp. 1325–1347, 1995.

[34] C. Scully, E. Castelli, E. Brearley, and M. Shirt, "Analysis and simulation of a speaker's aerodynamic and acoustic patterns for fricatives," *J. Phonetics*, vol. 20, pp. 39–51, 1992.

[35] C. Abry, P. Badin, and C. Scully, "Sound-to-gesture inversion in speech: The speech maps approach," in *Advanced Speech Applications* (K. Varghese, S. Pfleger, and J. Lefevre, eds.), pp. 182–196, Berlin: Springer Verlag, 1994.

[36] D. Beautemps, P. Badin, and R. Labiossiere, "Deriving vocal-tract area functions from midsagittal profiles and formant frequencies: A new model for vowels and fricative consonants based on experimental data," *Speech Communication*, vol. 16, pp. 27–47, 1995.

[37] P. Badin, D. Beautemps, R. Labiossiere, and J. Schwartz, "Recovery of vocal-tract geometry from formants for vowels and fricative consonants using a midsagittal-to-area function conversion model," *Journal of Phonetics*, vol. 23, pp. 221–229, 1995.

[38] M. M. Sondhi, "Estimation of vocal–tract areas: The need for acoustical measurements," *IEEE Transactions On Acoustics, Speech, and Signal Processing*, vol. ASSP–27, pp. 268–273, June 1979.

[39] L. Boë, P. Perrier, and G. Bailly, "The geometric vocal tract variables controlled for vowel production: proposals for constraining acoustic-to-articulatory inversion," *Journal of Phonetics*, vol. 20, pp. 27–38, 1992.

[40] A. Spanias, "Speech coding: A tutorial review," *Proceedings of the IEEE*, vol. 82, pp. 1541–1582, October 1994.

[41] J. L. Kelly and C. C. Lochbaum, "Speech synthesis," in *Proc. Fourth ICA*, p. Paper G42, 1962.

[42] S. Maeda, "Compensatory articulation during speech: evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model," in *Speech Production and Speech Modeling* (W. J. Hardcastle and A. Marchal, eds.), pp. 131–149, Kluwer Academic Publishers, 1990.

[43] P. Rubin, T. Baer, and P. Mermelstein, "An articulatory synthesizer for perceptual research," *Journal of the Acoustical Society of America*, vol. 70, pp. 321–328, August 1981.

[44] K. Ishizaka and J. L. Flanagan, "Synthesis of voiced sounds from a two mass model of the vocal cords," *Bell System Technical Journal*, vol. 50, pp. 1233–1268, July 1972.

[45] J. Schroeter, P. Meyer, and S. Parthasarathy, "Evaluation of improved articulatory codebooks and codebook access distance measures," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 393–396, April 1990.

[46] J. N. Larar, J. Schroeter, and M. M. Sondhi, "Vector quantization of the articulatory space," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, pp. 1812–1818, December 1988.

[47] K. N. Stevens, "Airflow and turbulence noise for fricative and stop consonants: Static considerations," *Journal of the Acoustical Society of America*, vol. 50, no. 4, pp. 1180–1192, 1971.

[48] P. Badin, C. Shadle, and T. Pham, "Frication and aspiration noise sources: contribution of experimental data to articulatory synthesis," in *International Conference on Spoken Language Processing*, vol. 1, (Yokohama, Japan), pp. 163–166, September 1994.

[49] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition.* PTR Prentice-Hall, Inc., 1993.

[50] B. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *Journal of the Acoustical Society of America*, vol. 24, pp. 175–184, 1952.

[51] J. P. Olive, A. Greenwood, and J. Coleman, *Acoustics of American English Speech.* New York: Springer–Verlag, 1993.

[52] P. Strevens, "Spectra of fricative noise in human speech," *Language and Speech*, vol. 3, pp. 32–49, 1960.

[53] G. Hughes and M. Halle, "Spectral properties of fricative consonants," *Journal of the Acoustical Society of America*, vol. 28, no. 2, pp. 303–310, 1956.

[54] J. Heinz and K. Stevens, "On the properties of voiceless fricative consonants," *Journal of the Acoustical Society of America*, vol. 33, pp. 589–596, May 1961.

[55] K. Harris, "Cues for the discrimination of American English fricatives in spoken syllables," *Language and Speech*, vol. 1, pp. 1–7, 1958.

[56] C. Shadle, "Modelling the noise source in voiced fricatives," in *15th International Congress on Acoustics*, (Trondheim, Norway), pp. 145–148, June 1995.

[57] C. Shadle, "The effect of geometry on source mechanisms of fricative consonants," *Journal of Phonetics*, vol. 19, pp. 409–424, 1991.

[58] C. Shadle, "Articulatory-acoustic relationships in fricative consonants," in *Speech Production and Speech Modelling* (W. Hardcastle and A. Marchal, eds.), pp. 187–209, Kluwer Academic, 1990.

[59] C. H. Shadle, *The Acoustics of Fricative Consonants.* PhD thesis, MIT, March 1985.

[60] H. Silverman and D. Morgan, "The application of dynamic programming to connected speech recognition," *IEEE ASSP Magazine*, vol. 7, pp. 6–25, July 1990.

[61] K. Shirai and M. Honda, "Estimation of articulatory motion from speech waves and its application for automatic recognition," in *Spoken Language Generation and Understanding* (J. Simon, ed.), pp. 87–99, D. Reidel Publishing Co., 1980.

[62] Q. Lin, G. Richard, J. Zou, D. Sinder, and J. Flanagan, "Use of tracttalk for adaptive voice mimic," in *Journal of the Acoustical Society of America*, (Washington, DC), Acoustical Society of America, May 1995. Presented at The 129th Meeting of the Acoustical Society of America.

[63] H. Yehia and F. Itakura, "Determination of human vocal-tract dynamic geometry from formant trajectories using spatial and temporal fourier analysis," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. I–477–I–480, 1994.

[64] B. Juang, L. R. Rabiner, and J. G. Wilpon, "On the use of bandpass filtering in speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-35, pp. 947–954, July 1987.